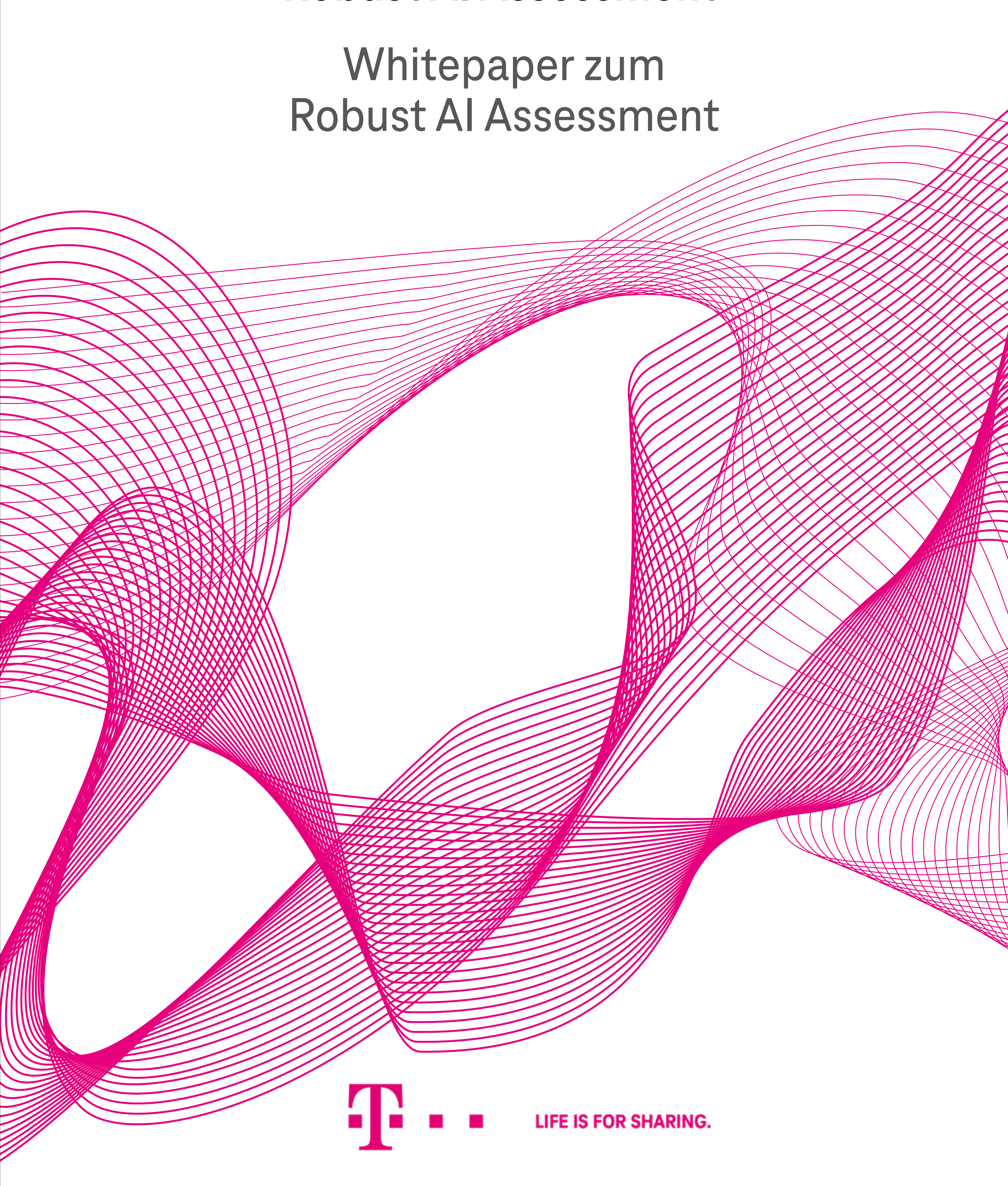
Whitepaper on Robust AI Assessment

Whitepaper zum Robust AI Assessment



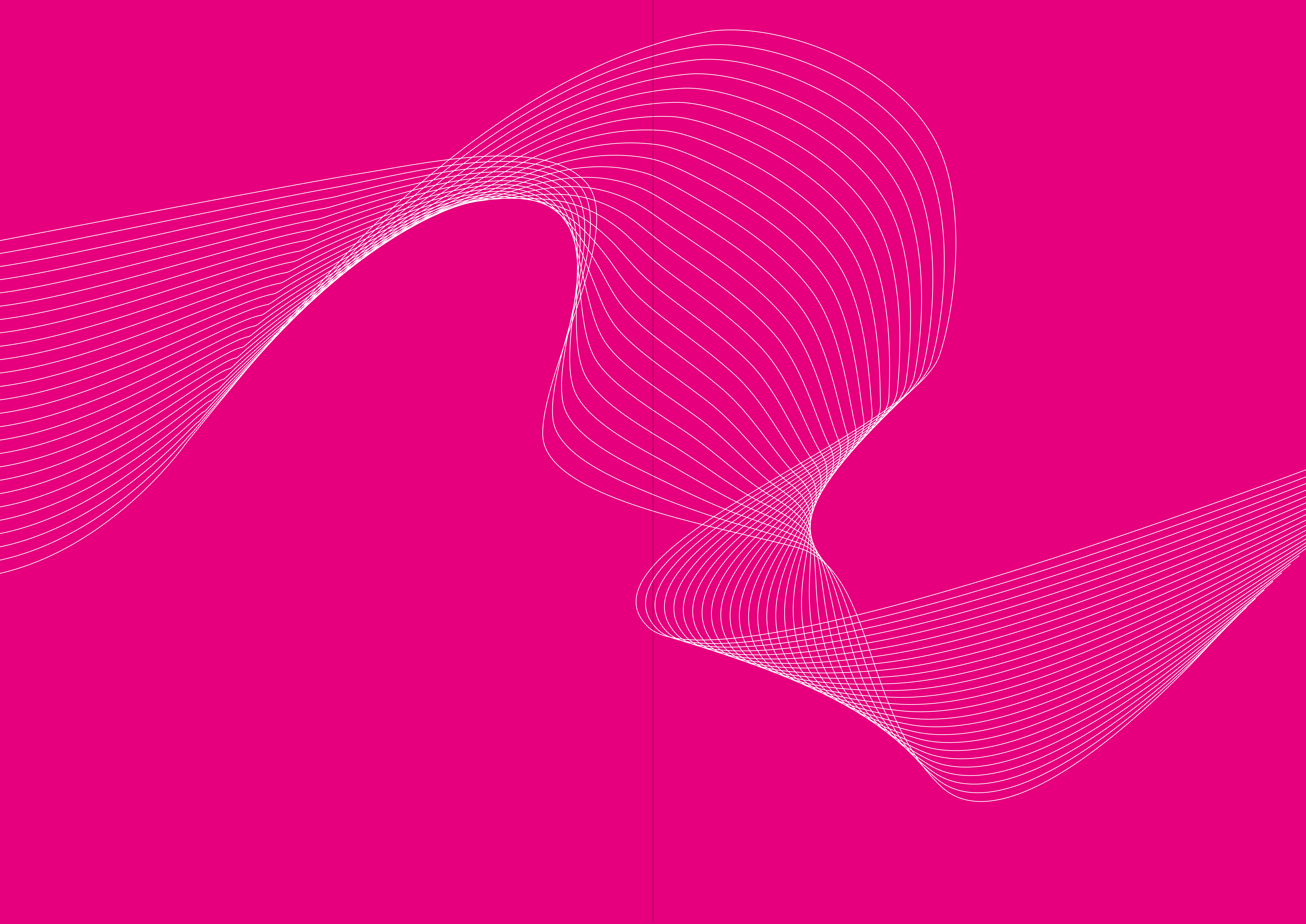


Table of Content

Inhaltsverzeichnis



Preface – Manuela Mackert

The social acceptance of artificial intelligence depends on its robustness, performance and responsible, sustainable use.

Artificial intelligence (AI) is already an integral part of our everyday lives. AI is increasingly "making" decisions that have a significant impact on our lives, including decisions involving healthcare, delivery, driving a car or granting loans.

Currently, AI models cannot yet question their own results when transferring what they have learned – they only apply trained patterns. The following scenario has already been successfully tested under laboratory conditions. Someone is sitting in a self-driving car and a drone projects traffic signs onto the road for fractions of a second, which the car heeds as it drives. This drone was programmed to cause an accident. From today's point of view, this would be an (almost) perfect crime, since there would be virtually no evidence left behind.[1]

AI has no feelings and no personality. For this reason, it is considered objective. It is solely guided by pure "logic", i.e. the algorithms. Whereas humans can be prejudiced and biased, AI decides rationally and impartially. But is this really the case? The following example arose at a large international corporation. The *AI Recruiting* Project not only produced completely unsuitable candidates, but also showed a bias against women. How was this possible? The algorithm can only learn from the data we provide, that is, from what is given to the machines as input. In the above recruiting example, the AI was trained with human recruitment decisions. In this example, it merely reproduced the existing biases displayed

Vorwort – Manuela Mackert

Die gesellschaftliche Akzeptanz von künstlicher Intelligenz hängt von der Robustheit, der Leistungsfähigkeit und dem verantwortungsvollen, nachhaltigen Umgang mit ihr ab.

Künstliche Intelligenz (KI) ist bereits heute aus unserem Alltag nicht mehr wegzudenken. KI trifft immer häufiger Entscheidungen, die unser Leben maßgeblich beeinflussen, wie z. B. in der Medizin, beim Autofahren oder bei der Kreditvergabe.

Aktuell können KI-Modelle bei der Übertragung von Gelerntem noch nicht die eigenen Ergebnisse hinterfragen – sie wenden lediglich trainierte Muster an. Folgendes Szenario wurde bereits unter Laborbedingungen erfolgreich getestet. Jemand sitzt in einem selbstfahrenden Auto und eine Drohne projiziert für Bruchteile einer Sekunde Verkehrsschilder auf die Straße, nach denen das Auto fährt. Diese Drohne wurde so programmiert, dass ein Unfall entstehen soll. Das wäre aus heutiger Sicht ein (nahezu) perfektes Verbrechen, da kaum nachvollziehbare Spuren hinterlassen werden.[1]

KI hat keine Gefühle und keine Persönlichkeit. Aus diesem Grund gilt sie als objektiv. Sie wird von reiner „Logik", den Algorithmen geleitet. Denn da, wo wir Menschen vorurteilsbelastet und voreingenommen sein können, entscheidet die KI absolut rational und unparteiisch – oder doch nicht? Das folgende Beispiel hat sich bei einem großen internationalen Konzern ereignet. Die KI schlug in einem Recruiting-Projekt nicht nur völlig ungeeignete Personen vor, sie war zudem voreingenommen gegenüber Frauen. Wie war das möglich? Der Algorithmus kann nur von unseren bereitgestellten Daten lernen, also davon, was den Maschinen als Dateninput gegeben wird. In diesem Beispiel wurde die KI mit menschlichen Einstellungsentscheidungen trainiert.

1 Cyber Security Labs @ Ben Gurion University, Phantom of the ADAS: Phantom Attacks on Driving Assistance Systems https://www.youtube.com/watch?v=1cSw4fXYqWI&feature=emb_logo. 2020.

by human decision-making, and therefore only suggested young, white, male candidates. There is therefore a risk that we transfer our individual biases through data into the fundamentals of the AI training, thus causing bias and stereotypes.

This means that we need highly reliable AI systems with the integrity to make timely and safe decisions in uncertain and unpredictable environments. In doing so, it is important to compensate for human weaknesses in decision-making. The AI systems must be resistant to targeted attacks and be built to process large amounts of data without abandoning established societal progress towards equity and equality.

Weaknesses in AI models have the potential to cause enormous damage to the company developing models. The security of the AI software system is therefore essential for future use. We must design the AI systems to be robust so that any negative consequences are minimized and that, despite everything, performance does not suffer.

Diese waren natürlich vorbelastet durch Voreingenommenheiten, welche die KI reproduzierte und so nur junge, weiße, männliche Personen vorschlug. Es besteht also die Gefahr, dass wir unsere individuelle Befangenheit über die Daten in die Trainingsgrundlagen der künstlichen Intelligenzen bringen und so darüber Voreingenommenheit sowie Stereotype manifestieren.

Das bedeutet im Umkehrschluss, dass wir hochgradig zuverlässige und integre KI-Systeme brauchen, die in unsicheren sowie unvorhersehbaren Umgebungen zeitnah und sicher Entscheidungen treffen können. Dabei gilt es, die menschlichen Schwächen bei Entscheidungen zu kompensieren. Die KI-Systeme müssen resistent gegen zielgerichtete Angriffe sein und dafür gebaut sein, große Datenmengen verarbeiten zu können, ohne etablierte zivilisatorische Errungenschaften aufzugeben.

Wenn Schwachstellen in KI-Modellen vorliegen, kann das Schadenspotenzial für das entwickelnde Unternehmen enorm sein. Die Sicherheit des KI-Softwaresystems ist somit essenziell für die zukünftige Nutzung. Wir müssen die KI-Systeme so robust gestalten, dass diese enormen Konsequenzen so gering wie möglich bleiben und trotz allem die Performance nicht leidet.

1.1. WHAT IS MEANT BY "ROBUST"?

Robust AI solutions are solutions that are impervious to influences and provide the expected performance in uncertain and unpredictable environments, without reproducing any human bias. When developing AI systems, these capabilities must be constructed and implemented with the utmost care.

For this reason, we have introduced a "Robust AI Assessment" at Deutsche Telekom (DT). This assessment is a component of our Digital Ethics Initiative to combine the development of the latest technologies with strong ethical standards. In this project, we focus on the analysis and evaluation of the robustness of AI systems.

1.2. RESPONSIBLE, SUSTAINABLE HANDLING (ACCOUNTABILITY) OF DECISIONS BY AI

The use of AI poses novel challenges to our existing value system, as well as the responsibility and liability associated with it. Currently, decisions and the associated responsibility can be attributed to individuals. How will this change in the future, when actions performed by AI software systems can or should no longer or only partially be controlled by humans? AI cannot be only a black box for us. We have to clearly define the systems and their processes. It is equally important to have complete documentation and assignment of responsibility.

1.1. WAS IST MIT „ROBUST“ GEMEINT?

Robuste KI-Lösungen sind Lösungen, die unempfindlich gegenüber Einflüssen sind und die in unsicheren sowie unvorhersehbaren Umgebungen die erwartete Performance bieten und dabei die Voreingenommenheit nicht reproduzieren. Bei der Entwicklung von KI-Systemen müssen diese Fähigkeiten gezielt umgesetzt werden.

Bei der Deutschen Telekom haben wir aus diesem Grund ein „Robust AI Assessment“ eingeführt. Dieses Assessment ist ein weiterer Baustein unserer Digital Ethics Initiative, um die Entwicklung von neuesten Technologien mit unseren ethischen Ansprüchen zu verbinden. Mit diesem Projekt fokussieren wir uns auf die Analyse und Bewertung der Robustheit von KI-Systemen.

1.2. VERANTWORTUNGSVOLLER, NACHHALTIGER UMGANG (ACCOUNTABILITY) MIT ENTSCHEIDUNGEN DURCH KI

Der Einsatz von KI stellt unser bisheriges Wertesystem sowie die damit verbundene Verantwortung und Haftung vor ganz neue Herausforderungen. Aktuell können Entscheidungen und die damit verbundene Verantwortung Personen zugerechnet werden. Wie wird sich das in der Zukunft ändern, wenn Handlungen die KI-Softwaresysteme durchführen, nicht mehr oder nur zum Teil von Menschen

Human beings must be able to intervene in human-machine interactions. For example, we must be able to stop or interrupt a system by means of an emergency stop switch. Many things can be digitalized or automated, but responsibility is not one of them. Day after day, more and more decisions are being made by AI software systems and thus delegated, but this does not mean that responsibility for the decisions made – possibly wrong decisions – also lies with the machines.

Digital transformation is to be understood holistically. It is not only about technology, but above all about how we enable *ethical leadership* – leadership that takes values, dignity, and rights of individuals and groups into account. From there, the question is then how to generate sustainable business guided by ethical leadership and innovative use of technology. In the foreseeable future, so-called intelligent technologies will complement human (emotional) intelligence. The emphasis here is on the word complement – we foresee a complex human-machine partnership, with its many ethical and business-relevant implications.

By introducing the "Robust AI Assessment," we want to transform social responsibility in the digital age and be a trustworthy partner for our customers.

kontrolliert werden können oder sollen? Die KI darf für uns keine Blackbox sein. Wir müssen die Systeme und deren Prozesse klar definieren. Ebenso wichtig ist eine lückenlose Dokumentation und Verantwortungszuordnung. Der Mensch muss bei Mensch-Maschine-Interaktionen die Möglichkeit haben, zu intervenieren oder aber ein System per Notausschalter anzuhalten oder zu unterbrechen. Es kann vieles digitalisiert oder automatisiert werden, aber Verantwortung gehört nicht dazu. Tag für Tag werden immer mehr Entscheidungen von KI-Softwaresystemen getroffen und somit delegiert, das heißt aber nicht, dass auch die Verantwortung für die getroffenen Entscheidungen – auch mögliche Fehlentscheidungen – bei den Maschinen liegt.

Digitale Transformation ist ganzheitlich zu verstehen. Es geht nicht nur um die Technologie, sondern vor allem auch darum, wie wir *Ethical Leadership*, d. h. Führung unter Berücksichtigung von Wertvorstellungen und der Würde und Rechte von Individuen und Gruppen befähigen und hieraus nachhaltiges Geschäft generiert werden kann. Die sogenannten intelligenten Technologien werden auf absehbare Zeit die menschliche (emotionale) Intelligenz ergänzen. Es geht um die Mensch-Maschine-Partnerschaft mit allen ethischen und business-relevanten Implikationen.

Mit dem Thema „Robust AI Assessment“ möchten wir der gesellschaftlichen Verantwortung im digitalen Zeitalter gerecht werden und unseren Kundinnen und Kunden eine vertrauensvolle Basis bieten.





Overview, Introduction & Theory

2.1. WHAT IS ROBUST AI?

2.1.1. WHAT IS AI?

The term Artificial Intelligence (AI) was first used in 1956 by the American computer scientist John McCarthy.[2] The term was used to describe systems that imitate human intelligence and develop their own methods to solve tasks independently – without human intervention – and that remains how it is defined today.[3]

Within DT and for the purposes of this whitepaper, projects are considered AI when they pursue the goal of implementing certain abilities of human thinking in computer systems, in order to enable the systems to solve tasks independently. The automation of business processes is not a new phenomenon. Until now, these processes have been defined directly by specific rules. Now, automation is increasingly implemented by AI systems that use "learned" rules derived from data. These stochastic AI systems are usually much more complex than previous systems were. While this makes it more difficult to understand decisions and actions, it also enables the solution of much more complex tasks.

The goal of automating processes has not changed for decades. The available means, however, have been greatly expanded in the past years.

Überblick, Einführung & Theorie

2.1. WAS IST ROBUSTE KI?

2.1.1. WAS VERSTEHEN WIR UNTER KI?

Der Begriff KI wurde 1956 erstmals durch den amerikanischen Computerwissenschaftler John McCarthy bekannt.[2] Bis heute wird die Bezeichnung KI genutzt, um Systeme zu beschreiben, die menschliche Intelligenz nachahmen und eigene Methoden entwickeln, um Aufgaben selbstständig – ohne menschlichen Eingriff – zu lösen.[3]

Innerhalb der Deutschen Telekom AG (DTAG) und in diesem Whitepaper werden alle Projekte als KI betrachtet, die das Ziel verfolgen, bestimmte Fähigkeiten des menschlichen Denkens in Computersystemen zu implementieren, sodass diese selbstständig Aufgaben lösen können. Die Automatisierung von geschäftlichen Prozessen ist kein neues Phänomen. Bisher wurden diese Prozesse direkt mittels bestimmter Regeln definiert. Nun wird Automatisierung zunehmend durch KI-Systeme umgesetzt, die auf „gelernten", aus Daten abgeleiteten Regeln basieren. Diese stochastischen KI-Systeme sind meist deutlich komplexer als es frühere Systeme waren. Dies führt zwar dazu, dass es schwieriger wird, Entscheidungen und Handlungen nachzuvollziehen, auf der anderen Seite ermöglicht es aber auch das Lösen von deutlich komplexeren Aufgaben.

Das Ziel möglichst automatisierter Prozesse hat sich seit Jahrzehnten nicht verändert. Nur die zur Verfügung stehenden Mittel sind andere geworden.

2 S. L. Andresen, "John McCarthy: father of AI," in IEEE Intelligent Systems, vol. 17, no. 5, pp. 84–85. DOI: 10.1109/MIS.2002.1039837. 2002.

3 Lucas, Bruce D., and Takeo Kanade. "An iterative image registration technique with an application to stereo vision." pp. 674. 1981.

2.1.2. DECEPTIONS FOR AI MODELS

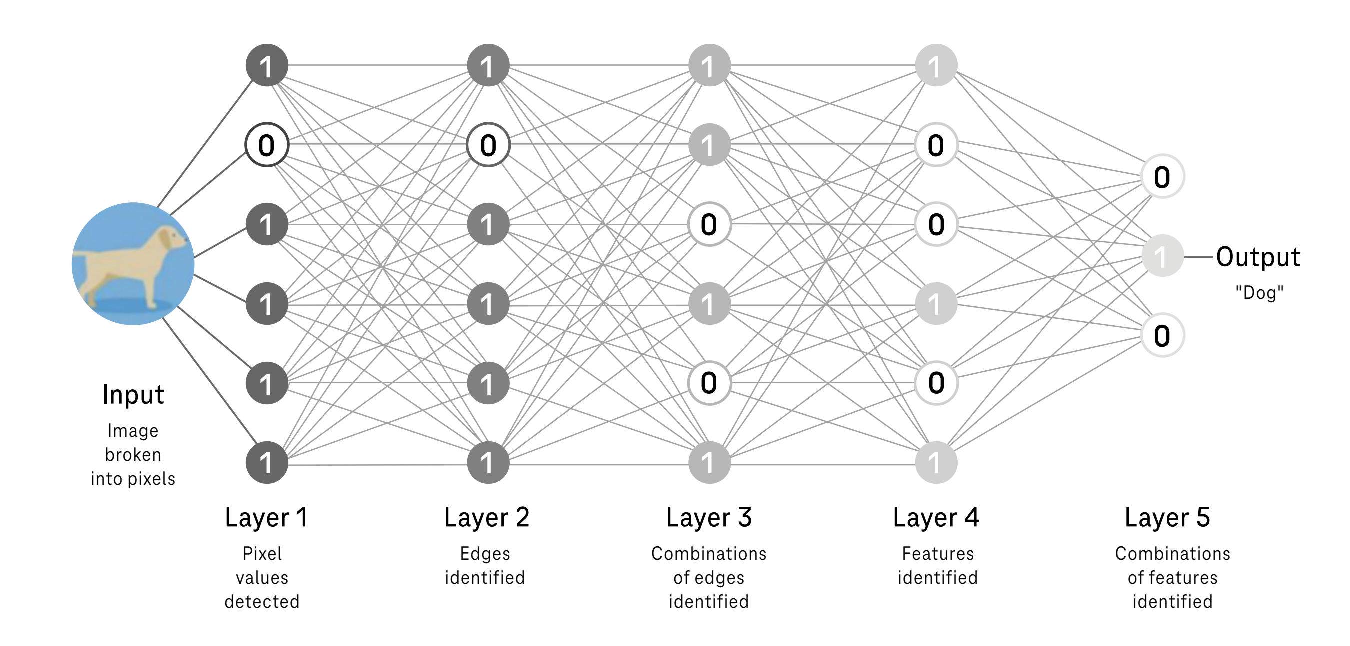
An AI model does not perceive reality as a whole. Instead, only previously trained patterns are recognized and processed. This can lead to effects that are comparable to traditional "human" optical illusions – a phenomenon that occurs when generalized thought patterns do not correspond to the actual situation.

While the effect in humans is usually minor and can be observed mainly in intentionally arranged shapes, for AI models, the effects are often far more significant. Borderline cases, where the AI model thinks it recognizes patterns that are not present in reality, quickly lead to wrong decisions. This is especially critical when it is not possible to understand how the system comes to a decision.[4]

2.1.2. TÄUSCHUNGEN FÜR KI-MODELLE

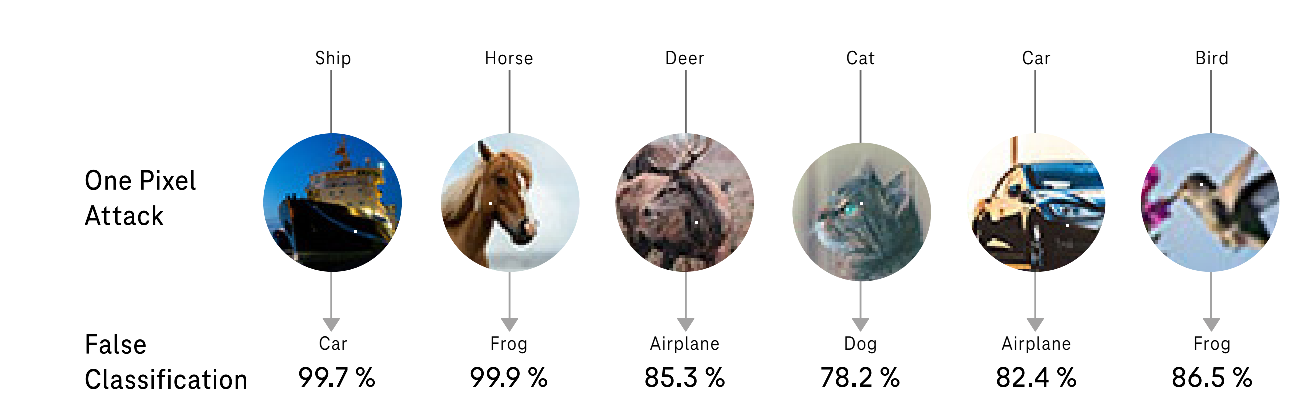
Ein KI-Model nimmt nicht die gesamte Wirklichkeit wahr. Stattdessen werden nur vorher trainierte Muster erkannt und weiterverarbeitet. Dadurch können Effekte auftreten, die vergleichbar mit „menschlichen" optischen Täuschungen sind. Ein Phänomen, das immer dann auftritt, wenn die generalisierten Denkmuster nicht mit der tatsächlich vorliegenden Situation übereinstimmen.

Während der Effekt beim Menschen meist unkritisch ist und sich hauptsächlich bei absichtlich angeordneten Formen beobachten lässt, sind die Auswirkungen für KI-Modelle oft größer. Grenzfälle, bei denen das KI-Modell meint, Muster zu erkennen, die in der Realität nicht vorhanden sind, führen hier schnell zu falschen Entscheidungen. Das ist besonders dann kritisch, wenn es nicht möglich ist nachzuvollziehen, wie das System zu einer Entscheidung kommt.[4]



4 Gomez-Villa, Alexander, et al. "Convolutional neural networks can be deceived by visual illusions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

Figure: https://www.quantamagazine.org/machine-learning-confronts-the-elephant-in-the-room-20180920



AI models distinguish between two types of borderline cases:

1. Natural perturbations: *Natural perturbations* are deviations that can occur "naturally" in the input data. These natural patterns can be caused by natural occurrences such as rain on images or noises in the background of audio recordings.[5]

2. Adversarial attack: *Adversarial attacks* are attacks that are carried out using data deliberately created with the aim of disrupting an AI model. In order to generate these purposefully produced inputs, the attacker trains their own AI model (so-called adversarial model) to modify inputs in such a way that they are as difficult as possible for the original model to process. In most cases, the changes are not recognizable for a human being, while for the machine the input changes completely.[6]

For an AI-based system to be robust, the results of the system should be influenced as little as possible by both *natural perturbations* and *adversarial attacks*.

Bei KI-Modellen unterscheidet man zwischen zwei Arten von Grenzfällen:

1. Natural Perturbations: Als *Natural Perturbations* werden Abweichungen bezeichnet, die „natürlich" in den Eingabedaten entstehen können. Diese natürlichen Muster können beispielsweise durch Regen auf Bildern oder durch Rauschen und andere Geräusche im Hintergrund von Audioaufnahmen entstehen.[5]

2. Adversarial Attack: Als *Adversarial Attack* bezeichnet man Angriffe, die mithilfe von bewusst erstellten Daten durchgeführt werden, mit dem Ziel ein KI-Modell zu stören. Um diese gezielt hergestellten Eingaben zu generieren, trainieren Angreifende das eigene KI-Modell (*sog. Adversarial Model*) darauf, Eingaben so zu modifizieren, dass sie möglichst schwer für das ursprüngliche Modell zu verarbeiten sind. Meist sind die Änderungen für einen Menschen nicht erkennbar, während sich für die Maschine der Input komplett ändert.[6]

Damit ein KI-basiertes System robust ist, sollten die Ergebnisse des Systems so wenig wie möglich durch beide Arten von Grenzfällen beeinflusst werden.

5 Ozdag, Mesut, et al. On the Susceptibility of Deep Neural Networks to Natural Perturbations. Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States). 2019.

Figure: Prof. Dr. Ing. Marco Huber, Fraunhofer IPA

6 Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572. 2014.

Whether it is more important to react to natural influences or targeted attacks depends strongly on the environment and the respective application.

2.1.3. BIAS AND AI MODELS (BIAS)

In contrast to classical development methods, the result of an AI system is not explicitly defined by specific rules. Instead, the system is trained with the help of optimization algorithms in such a way that regularities within a database are mapped. AI systems therefore do not automatically map reality, but only model based on a given database. Whether an AI decision is actually based on reality, or whether it follows an incorrect or distorted representation of reality, depends heavily on the data used to train the model. Discriminatory or prejudicial AI models, which have been repeatedly discussed, are therefore by no means caused by the AI models themselves. However, widespread social bias often influences the data collection in samples or the interpretation; this leads to biased datasets. The bias implicit in the data is thus reproduced by AI models.[7]

To reduce bias in AI models, training data should be thoroughly checked for possible distortions so that these can be eliminated before the training process. Importantly, bias is often implicit in the data. In many cases, it is not enough to remove ethically questionable attributes such as ethnicity or religious affiliation, as these can often be reconstructed using a combination of other, initially harmless attributes such as age, place of residence and creditworthiness. For example, it is possible that a person's national origin indirectly has a strong influence on the decision of a data-based AI model, even though the origin was never explicitly included in the training data. To exclude discrimination of this kind, fairness

Ob es wichtiger ist, auf natürliche Einflüsse oder zielgerichtete Angriffe zu reagieren, hängt dabei stark vom Einsatzumfeld und dem jeweiligen Anwendungsfall ab.

2.1.3. VOREINGENOMMENHEIT UND KI-MODELLE (BIAS)

Im Gegensatz zu klassischen Entwicklungsmethoden wird das Ergebnis eines KI-Systems nicht explizit anhand bestimmter Regeln festgeschrieben. Stattdessen wird das System mithilfe von Optimierungsalgorithmen so trainiert, dass Gesetzmäßigkeiten innerhalb einer Datenbasis abgebildet werden. KI-Systeme bilden also nicht automatisch die Wirklichkeit ab, sondern modellieren lediglich anhand einer vorgegebenen Datengrundlage. Ob eine KI-Entscheidung dabei tatsächlich in der Wirklichkeit begründet ist oder einer falschen oder verzerrten Darstellung der Wirklichkeit folgt, hängt stark von der Datengrundlage ab, die für das Training des Modells genutzt wird. Der Grund für diskriminierende oder vorverurteilende KI-Modelle, die immer wieder thematisiert werden, liegt also keinesfalls in den KI-Modellen selbst. Gesellschaftlich verbreitete Voreingenommenheit beeinflusst aber oft die Datenerhebung bei Stichproben oder die Interpretation und führt so zu einer vorbelasteten Datengrundlage. Die implizit in den Daten enthaltene Voreingenommenheit wird also nur durch die KI-Modelle reproduziert.[7]

Trainingsdaten sollten daher unbedingt gründlich auf mögliche Verzerrungen überprüft werden, damit diese vor dem Trainingsprozess behoben werden können. Wichtig ist, dass Voreingenommenheit oft implizit in den Daten enthalten ist. Es reicht nicht, ethisch fragwürdige Attribute wie beispielsweise Ethnie oder Religionszugehörigkeit zu entfernen, da diese oft über eine Kombination von anderen, zunächst harmlos scheinenden Attributen wie Alter, Wohnort und Kreditwürdigkeit wieder rekonstruiert werden können. So ist es beispielsweise möglich, dass die

7 Du, Mengnan, et al. "Fairness in deep learning: A computational perspective." IEEE Intelligent Systems. 2020.

metrics can be used to technically analyze how ethically relevant attributes are represented in the data.[8]

In addition to the analysis of the training data, the patterns describing why the AI model chooses a particular outcome should also be analyzed. Here it is important to check whether the decision-making process is justifiable from an ethical point of view and fulfils basic moral principles. A variety of methods can be used to investigate this. The options range from traditional sensitivity analyses, which examine what effect a certain change in an attribute has on the model's decision[9], to more modern approaches such as SHAP-Values-Analysis, a game theory method that evaluates the importance of attributes.[10]

Through the interaction of a truly representative database and verified, robust model behavior, prejudice and discrimination by AI systems can be prevented in the long term.

2.1.4. DEFINING ROBUST AI

In computer science, systems are described as "robust" if they function fault-tolerantly despite adverse conditions. While this superordinate term often refers to the prevention of system crashes and thus to the failure rate, robust AI considers the process of decision-making. Traditional IT security topics such as system availability are only dealt with peripherally.

Herkunft einer Person indirekt einen starken Einfluss auf die Entscheidung eines datenbasierten KI-Modells hat, obwohl die Herkunft nie explizit in den Trainingsdaten enthalten war. Um Diskriminierung dieser Art auszuschließen, können Fairness-Metriken genutzt werden, die technisch analysieren, wie ethisch-relevante Attribute in den Daten vertreten sind.[8]

Zusätzlich zu der Analyse der Trainingsdaten sollten auch die Muster analysiert werden, die beschreiben, weshalb das KI-Modell sich für ein bestimmtes Ergebnis entscheidet. Hier gilt es zu überprüfen, ob die Entscheidungsfindung unter ethischen Gesichtspunkten vertretbar ist und die moralischen Grundprinzipien erfüllt. Um dies zu untersuchen, kann eine Vielzahl an Methoden genutzt werden. Die Optionen reichen von herkömmlichen Sensitivitätsanalysen, die untersuchen, welche Auswirkung eine bestimmte Änderung eines Attributs auf die Entscheidung des Modells hat[9], bis hin zu moderneren Ansätzen wie die SHAP-Values-Analyse, welche die Wichtigkeit der Attribute über eine Methode aus der Spieltheorie bewertet.[10]

Durch das Zusammenspiel einer tatsächlich repräsentativen Datenbasis und einem verifizierten, robusten Modellverhalten können Vorverurteilung und Diskriminierung durch KI-Systeme nachhaltig verhindert werden.

2.1.4. DEFINITION „ROBUSTE KI“

In der Informatik werden Systeme als „robust“ bezeichnet, die trotz widriger Bedingungen fehlertolerant funktionieren. Während bei diesem übergeordneten Begriff oft

8 Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

9 Féraud, Raphael, and Fabrice Clérot. "A methodology to explain neural network classification." Neural networks pp. 237–246. 15.2.2002.

10 Antwarg, Liat, Bracha Shapira, and Lior Rokach. "Explaining anomalies detected by autoencoders using SHAP." arXiv preprint arXiv:1903.02407. 2019.

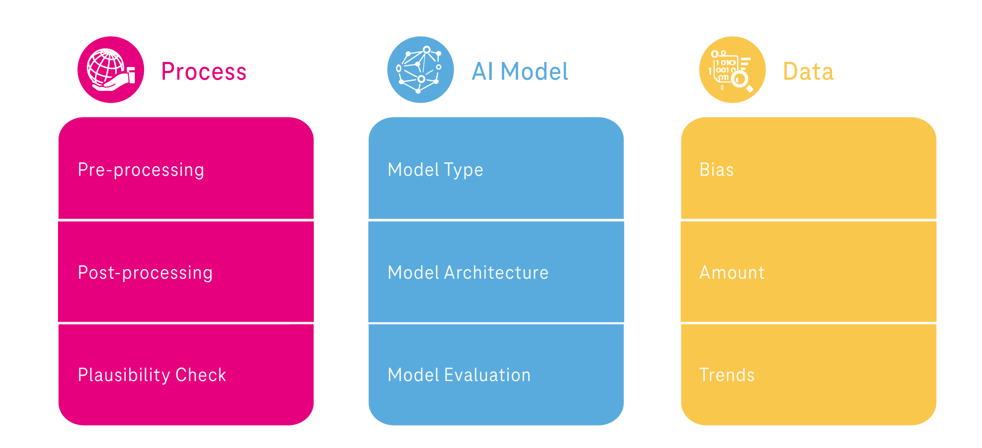
In general, robust AI can be defined as follows: A robust AI is an unbiased, AI-based system that makes technically sound decisions that are not biased and whose decision-making process is not significantly influenced by *natural perturbations* or *adversarial attacks*. The decisions of a robust AI system reflect reality and are based on data that is as free of bias as possible.[11]

Robust AI is divided into three sub-domains and depends mainly on three factors: (1) the robustness of the (business) process in which the system is integrated, (2) the structure and type of the AI model used, and (3) the database that is the basis for training the model.

das Verhindern von Systemabstürzen gemeint ist und sich damit auf die Ausfallrate bezieht, betrachtet robuste KI den Prozess der Entscheidungsfindung. Dabei werden herkömmliche IT-Sicherheitsthemen wie Verfügbarkeit des Systems nur peripher behandelt.

Im Allgemeinen kann robuste KI wie folgt definiert werden: Eine robuste KI ist ein unvoreingenommenes, KI-basiertes System, das technisch nachvollziehbare Entscheidungen trifft, die nicht durch Voreingenommenheit belastet sind und dessen Entscheidungsfindungsprozess durch *Natural Perturbations oder Adversarial Attacks nicht signi*fikant beeinflusst wird. Die Entscheidungen eines robusten KI-Systems spiegeln die Wirklichkeit wider und basieren auf einer möglichst verzerrungsfreien Datengrundlage.[11]

Robuste KI teilt sich in drei Unterdomänen auf und ist maßgeblich von drei Faktoren abhängig:
(1) der Robustheit des (Geschäfts-) Prozesses, in den das System eingebunden ist, (2) dem Aufbau und der Art des genutzten KI-Modells sowie (3) der Datenbasis, welche die Grundlage für das Training des Modells ist.



11 Mirman, Matthew, Timon Gehr, and Martin Vechev. "Differentiable abstract interpretation for provably robust neural networks." International Conference on Machine Learning. 2018.

2.2. WHY IS IT IMPORTANT TO US THAT OUR AI SOLUTIONS ARE ROBUST?

AI-based systems are known to provide accurate and high-quality predictions, especially if a good database is available. However, the more complex the data models an AI represents, the more difficult it becomes to understand why the system chooses a certain outcome and whether the decision is actually justified. Therefore, robust AI is a necessary prerequisite for a trustworthy AI. This is especially necessary given that AI technology will increasingly be used for decisions with high risk potential; it must be possible to understand how a decision is made. It is particularly important to check whether a decision is resilient or whether it is based on a misunderstanding of artificial intelligence in order to meet our responsibility for fair and ethical action.

The Robust AI Assessment for AI is part of the overall efforts of DT to combine the latest technologies with ethical standards. The project is part of the Robust AI program, which focuses on analyzing and evaluating the robustness of AI models. In a cooperation with experts from Ben-Gurion University in Israel and the German start-up Neurocat, which specializes in robust AI, the AI experts at Telekom Innovation Laboratories are engaged in measuring and technically analyzing the robustness of internal and external AI-based products and services of DTAG and thus identifying potential for improvement. The goal of ethical technology development occupies DT along the entire value chain. Companies that significantly shape the future of society through the

2.2. WARUM IST ES UNS WICHTIG, DASS UNSERE KI-LÖSUNGEN ROBUST SIND?

KI-basierte Systeme sind dafür bekannt – besonders, wenn eine gute Datenbasis vorhanden ist – akkurate und qualitativ sehr hochwertige Vorhersagen zu treffen. Doch je komplexer die Datenmodelle werden, die eine KI abbildet, desto schwieriger wird es zu verstehen, aus welchem Grund sich das System für ein bestimmtes Ergebnis entscheidet und ob die Entscheidung tatsächlich gerechtfertigt ist. Deswegen ist robuste KI eine notwendige Voraussetzung für eine vertrauenswürdige KI. Gerade im Hinblick darauf, dass KI-Technologie zunehmend auch für Entscheidungen mit hohem Risikopotenzial eingesetzt werden wird, muss es möglich sein, nachzuvollziehen, wie eine Entscheidung zustande kommt. Dabei ist es insbesondere wichtig zu prüfen, ob eine Entscheidung belastbar ist oder ob sie auf einem Missverständnis der künstlichen Intelligenz beruht, um unserer Verantwortung für faires und ethisches Handeln gerecht zu werden.

Das Robust AI Assessment für KI ist Teil der übergreifenden Bestrebungen der Deutschen Telekom AG (DTAG), neueste Technologien mit ethischen Ansprüchen zu verbinden. Das Projekt ist ein Teil des Robust-AI-Programms, das sich insgesamt darauf fokussiert, die Robustheit von KI-Modellen zu analysieren und zu bewerten. In einer Kooperation mit der Ben-Gurion-Universität in Israel und dem deutschen Start-Up Neurocat, das sich auf robuste KI spezialisiert hat, beschäftigen sich die KI-Expertinnen und -Experten der Telekom Innovation Laboratories damit, die Robustheit von internen und externen KI-basierten Produkten und Dienstleistungen der DTAG zu messen, technisch zu analysieren und so Verbesserungspotenziale

development of new, disruptive technologies and offerings bear an immense responsibility. In order to live up to this responsibility, DT has developed its own "Digital Ethics" guidelines that address the responsible and ethical use of artificial intelligence and establish self-imposed rules. *The detailed version of these guidelines can be found on the Group website.*

The Robust AI Assessment, which DT wants to use to integrate ethical aspects into the development process of new AI systems, therefore fits seamlessly into a series of approaches that promote a future in which companies meet their social responsibility in the digital age.

zu identifizieren. Das Ziel ethischer Technologieentwicklung beschäftigt die Deutsche Telekom über die gesamte Wertschöpfungskette hinweg. Unternehmen, die durch die Entwicklung neuer, disruptiver Technologien und Angebote die Zukunft der Gesellschaft maßgeblich prägen, tragen eine immense Verantwortung. Um dieser Verantwortung gerecht zu werden, hat die Deutsche Telekom eigene „Digital Ethics“ Leitlinien entwickelt, die den verantwortungsvollen und ethischen Umgang mit künstlicher Intelligenz thematisieren und selbstverpflichtende Regeln aufstellen. *Die ausführliche Version dieser Leitlinien ist auf der Konzernwebseite zu finden.*

Das Robust AI Assessment, das die Deutsche Telekom dafür nutzen möchte, um ethische Aspekte in den Entwicklungsprozess neuer KI-Systeme einzubinden, fügt sich also nahtlos in eine Reihe von Ansätzen ein, die eine Zukunft fördern, in der Unternehmen ihrer gesellschaftlichen Verantwortung im digitalen Zeitalter gerecht werden.

2.3. HOW ROBUST SHOULD AI SOLUTIONS BE?

In general, the more robust an AI-based system is, the better. But is there a point where increasing robustness is no longer helpful, maybe even counterproductive?

How robust a system should actually be depends heavily on the application. An intelligent vending machine naturally has different requirements than an AI system of a critical infrastructure system. AI models are often used as expert systems that are specialized to perform a certain complex task in a narrowly defined application area.
The underlying data models are very good at deriving and learning specific regularities under certain conditions from a data set. Increasing the robustness of an AI system means extending the range in which the model can make plausible decisions. In extreme cases, the AI system changes from an expert to a generalist and thus loses the ability to solve the complex, special task for which it was trained. Since the complexity of the model is generally limited, for example, by the computing power required for training and the amount of data available, decisions can no longer be made with the same nuance as by a pure expert system.[12]

In general, the robustness of an AI system is increased by means of use case-specific measures that are often a tradeoff between model robustness, predictive performance and training effort. The development of further measures to increase the robustness causes a high level of effort, but does not create any added value. More measures than necessary do not make the model more robust but may even lead to worse decisions.[13]

AI models should therefore be, at best, somewhat more robust than their intended use requires in order to compensate for possible misjudgments in the assessment. However, it is not recommended to increase the robustness of such a system beyond what is necessary, as the performance of the AI model tends to deteriorate while the development effort increases.

2.3. WIE ROBUST SOLLTEN KI-LÖSUNGEN SEIN?

Generell gilt: Je robuster ein KI-basiertes System ist, desto besser. Doch gibt es einen Punkt, an dem die Steigerung der Robustheit nicht mehr weiterhilft, vielleicht sogar kontraproduktiv ist?

Wie robust ein System tatsächlich sein sollte, ist stark vom Anwendungsfall abhängig. Ein intelligenter Getränkeautomat hat natürlich andere Anforderungen als ein KI-System einer kritischen Infrastruktur. KI-Modelle werden oft als Systeme eingesetzt, die darauf spezialisiert sind, eine bestimmte komplexe Aufgabe in einem eng definierten Anwendungsbereich zu erledigen.
Die zugrundeliegenden Datenmodelle sind sehr gut darin, spezifische Gesetzmäßigkeiten unter bestimmten Bedingungen aus einer Datengrundlage abzuleiten und zu lernen. Wenn man die Robustheit eines KI-Systems steigert, bedeutet dies, den Bereich zu erweitern, in dem das Modell plausible Entscheidungen treffen kann. Im Extremfall wandelt sich das KI-System dabei vom Experten zum Generalisten und verliert damit die Fähigkeit, die komplexe, spezielle Aufgabe zu lösen, für die es trainiert wurde. Da die Komplexität des Modells generell beispielsweise durch die benötigte Rechenleistung für das Training und die verfügbare Datenmenge begrenzt ist, können die Entscheidungen nicht mehr so nuanciert getroffen werden, wie es ein reines Expertensystem könnte.[12]

Allgemein wird die Robustheit eines KI-Systems mithilfe von gezielt angewendeten Maßnahmen gesteigert. Das Entwickeln von weiteren Maßnahmen zur Steigerung der Robustheit verursacht einen hohen Aufwand, der aber keinen Mehrwert schafft. Mehr Maßnahmen als nötig machen das Modell nicht robuster, sondern können sogar zu schlechteren Entscheidungen führen.[13]

KI-Modelle sollten deswegen im besten Fall etwas robuster sein als ihr Einsatzzweck erfordert, um eventuelle Fehleinschätzungen bei der Beurteilung auszugleichen. Es ist aber nicht empfehlenswert, die Robustheit eines solchen Systems in großem Maße über die Anforderungen hinaus zu steigern, da dabei tendenziell die Performance des KI-Modells nachlässt, obwohl der Entwicklungsaufwand steigt.

12 Valiant, Leslie G. "Knowledge infusion: In pursuit of robustness in artificial intelligence." IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2008.

13 Anderson, Greg, et al. "Optimization and abstraction: A synergistic approach for analyzing neural network robustness." Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation. 2019.



Robustness Self-Assessment

3.1. WHY DO WE USE A SELF-ASSESSMENT FRAMEWORK?

Within DT there are currently a large number of active projects and systems that are developing and/or using artificial intelligence. The number of departments involved in AI development is constantly increasing. The goal must be to give technical project managers, as well as AI experts and data scientists, an overview of which robustness the use case requires and whether the measures currently planned or developed are sufficient. This should be done as early as possible in the development phase of a new AI project.

The earlier the topic of robustness is considered, the better the measures can be adapted to the respective application. In contrast to conventional processes, in which measures are often identified and developed at a late stage, this "Robustness by Design" approach creates systems that are inherently robust and cannot be implemented only by means of applied measures.[14]

To ensure that our AI-based products are not vulnerable to attacks and other variations in input data, we also use technical analysis that a new AI-based application must pass before it can be used. As previously mentioned, the Robust AI program is currently developing and improving technical analysis

Robustness Self-Assessment

3.1. WARUM NUTZEN WIR EIN SELF-ASSESSMENT FRAMEWORK?

Innerhalb der Deutschen Telekom AG gibt es aktuell eine Vielzahl an aktiven Projekten und Systemen, die künstliche Intelligenz entwickeln und/oder einsetzen. Die Anzahl der Abteilungen, die an der KI-Entwicklung beteiligt sind, steigt ständig. Ziel muss es sein, sowohl technischen Projektleitungen als auch KI-Expertinnen und -Experten sowie Data Scientists frühzeitig im Entwicklungsprozess eines neuen KI-Projektes einen Überblick zu geben, welche Robustheit der Anwendungsfall erfordert und ob die Maßnahmen, die aktuell geplant oder entwickelt werden, ausreichend sind.

Je früher innerhalb der Entwicklung eines neuen KI-basierten Tools das Thema Robustheit betrachtet wird, desto besser können die Maßnahmen auf den jeweiligen Anwendungsfall abgestimmt werden. Anders als in herkömmlichen Prozessen, bei denen Maßnahmen oft erst spät identifiziert und entwickelt werden, entstehen durch dieses „Robustness by Design" Systeme, die inhärent robust sind und nicht erst durch aufgesetzte Maßnahmen einsetzbar werden.[14]

Damit wir sicherstellen können, dass unsere KI-basierten Produkte tatsächlich nicht für Angriffe und andere Abweichungen der Input-Daten anfällig sind, setzen wir zusätzlich technische Analysen ein, die eine neue KI-basierte Anwendung vor dem Einsatz bestehen muss. Wie zuvor erwähnt, werden derzeit innerhalb des Robust-AI-Programmes neben dem Self-Assessment-Konzept ebenfalls

14 Valiant, Leslie G. "Knowledge infusion: In pursuit of robustness in artificial intelligence." IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2008.

options in addition to the self-assessment concept. Ideally, the technical analysis is mainly used to verify the proper implementation and functionality of the selected measures, as well as to automatically test new releases.

3.2. HOW IS THE SELF-ASSESSMENT CONCEPT STRUCTURED?

The self-assessment concept is based on the principle of a gap analysis. The first part considers the robustness requirements that are directly related to the use case and thus representative of the target specification. The second part assists in developing the planned or implemented measures and thus represents the actual state. The analysis values of both parts can then be compared in order to identify differences between target and actual state.

technische Analysemöglichkeiten weiterentwickelt und verbessert. Im Idealfall wird die technische Analyse hauptsächlich dafür genutzt, die ordnungsgemäße Umsetzung und Funktionalität der gewählten Maßnahmen zu verifizieren sowie neue Releases automatisiert zu prüfen.

3.2. WIE IST DAS SELF-ASSESSMENT-KONZEPT STRUKTURIERT?

Das Self-Assessment-Konzept basiert auf dem Prinzip einer Gap-Analyse. Der erste Teil befasst sich mit den Robustheitsanforderungen, die unmittelbar durch den Anwendungsfall bedingt sind und bildet damit die Soll-Vorgabe ab. Der zweite Teil behandelt die geplanten bzw. umgesetzten Maßnahmen und steht somit für den Ist-Zustand. Die Analysewerte beider Teile können dann gegenübergestellt werden, um Differenzen zwischen Soll- und Ist-Zustand zu identifizieren.

Self-Assessment Tool

Addressed user group

Data Scientist
Product Owner

Addressed solutions

POC
Detailed concept
Implemented solution (early stage)

Implemented solution
Implemented multiple technical robustness measures

Use Case Details

Process
AI Model
Data

First Step.
1 Assessing the required level of robustness.

Second Step.
2 Assessing the actual level of robustness.

Outcome

Required and actual level of rebustness
Shaping AI Robustness awareness
Recommen-dation

Technical Assessment

aidkit Framework. Testing against adversarial attacks and natural perturbations.

The self-assessment framework is comprised of 35 weighted questions. Approximately one third of these belong to the first part and relate to the possible effects of a malfunction of the model under consideration, the probability of an attack or the danger of unexpected input. The analysis of the actual state consists of 24 questions, which are divided into sub-sections based on the domains of robust AI: (1) robust process, (2) robust AI model, and (3) robust data basis.

3.3. WEIGHTING OF THE QUESTIONS

Since each question has an distinctly different influence on the robustness of an AI solution, a weighting system is used to put the questions into relation. All questions in each category were weighted and then evaluated using the "*paired comparisons*" method. The "*paired comparisons*" method compared questions in pairs and assigns these a factor that describes the relative value between the questions ("A good answer to question 1 is worth twice as much as a good answer to question 2"). In the first step, no attention was paid to keeping all ratings consistent across multiple question pairs, as this promotes the most accurate assessment possible. In the second step, all conflicts were then listed and resolved one after the other by adjusting the ratings. The resulting weightings were critically reviewed in expert and user interviews and were slightly adjusted for the final version of the assessment. To ensure that the *Robustness Score* result values are within the desired value range, the resulting weights were uniformly scaled.

Insgesamt umfasst das Self-Assessment Framework 35 gewichtete Fragen. Ungefähr ein Drittel davon zählen zum ersten Teil und beziehen sich auf die möglichen Auswirkungen einer Fehlfunktion des betrachteten Modells, die Wahrscheinlichkeit eines Angriffes oder die Gefahr einer unerwarteten Eingabe. Die Analyse des Ist-Zustands besteht aus 24 Fragen, die anhand der Domänen robuster KI in Teilbereiche aufgeteilt sind: (1) robuster Prozess, (2) robustes KI-Modell und (3) robuste Datengrundlage.

3.3. GEWICHTUNG DER FRAGEN

Da jede Frage einen individuell unterschiedlichen Einfluss auf die Robustheit einer KI-Lösung hat, wird ein Gewichtungssystem genutzt, um die Fragen ins Verhältnis zu setzen. Dabei wurden zuerst alle Fragen der einzelnen Kategorien gewichtet und anschließend anhand der *„Paired-Comparison-Methode"* bewertet. Diese Methode vergleicht Fragenpaare miteinander, die dadurch einen Faktor erhalten, der den relativen Wert zwischen den Fragen beschreibt („Eine gute Antwort auf Frage 1 ist doppelt so viel wert wie eine gute Antwort auf Frage 2."). Im ersten Schritt wurde explizit nicht darauf geachtet, alle Bewertungen über mehrere Fragenpaare konsistent zu halten, da so eine möglichst genaue Einschätzung gefördert wird. Im zweiten Schritt wurden dann alle Konflikte aufgeführt und nacheinander durch das Anpassen der Bewertungen aufgelöst. Die dabei entstandenen Gewichtungen wurden in Interviews mit Expertinnen und Experten sowie Nutzerinnen und Nutzern kritisch hinterfragt und für die finale Version des Assessments leicht angepasst. Damit die *Robustness-Score*-Ergebniswerte im gewünschten Wertebereich liegen, wurden die entstandenen Gewichte uniform skaliert.

3.4. WHAT VALUES DOES THE ROBUST AI ASSESSMENT PROVIDE?

The results of the Robust AI Assessment consist of two values: (1) the *Required Robustness Score* and (2) the *Actual Robustness Score*. These two values, which express the required robustness and the implemented robustness, can be directly compared. As a result, possible gaps can be detected at a glance.

The evaluation of the Robust AI Assessment is based on an additive concept. This means that the *basic score* starts at zero and can be easily increased by any valid measure. The selection of this *scoring system*, *which* gives a value of 0–5 in each category, is based on the assumption that models that are not trained to act robustly do not become robust by chance. Current methods for training AI models have usually been developed to find the simplest, most pronounced patterns in data sets. These are usually not the most robust patterns. In addition, finding more robust patterns usually requires more data and a longer development phase to adapt the model to the data conditions.

Analogous to the weighting of the questions, a factor between -1 and +2 was assigned to the possible answers of the questions, which is represented during the assessment by one of the following symbols: -, o, +, ++. This factor is multiplied by the weighting of the respective question and thus flows weighted into the overall result. For the assessment of the required robustness, answers that indicate a high required robustness receive a high factor. For the assessment of the implemented measures, answers that indicate the implementation of valid measures also receive a high factor.

To obtain the two result values – the *Required Robustness Score* and the *Actual Robustness Score* – between 0 and 5, all question results are summed up under the respective category.

3.4. WIE SETZEN SICH DIE ERGEBNISWERTE ZUSAMMEN?

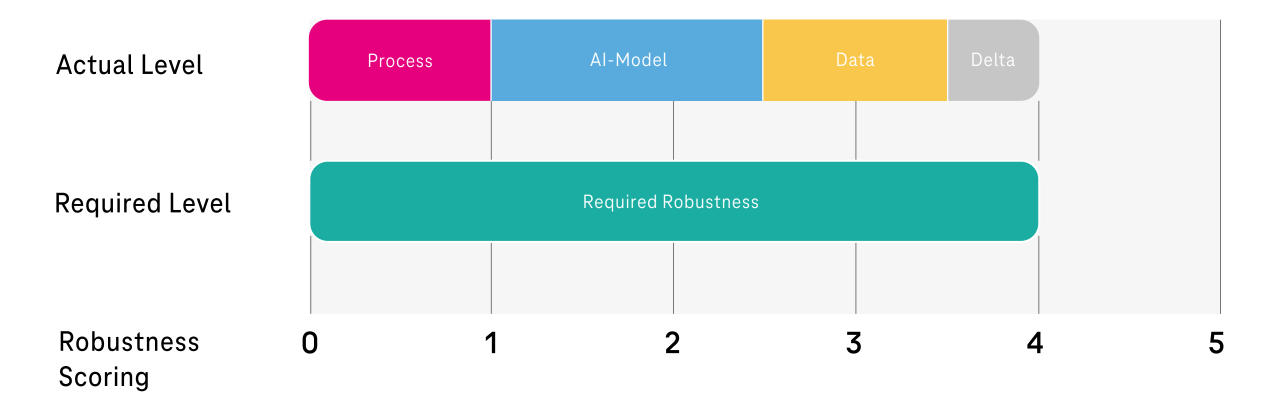
Das Ergebnis des Robust AI Assessments besteht aus zwei Werten: (1) dem *Required Robustness Score* und (2) dem *Actual Robustness Score*. Diese beiden Werte, welche die benötigte Robustheit und die umgesetzte Robustheit ausdrücken, können direkt gegenübergestellt werden. So können mögliche Gaps auf einen Blick erkannt werden.

Die Wertung des Robust AI Assessments basiert auf einem additiven Konzept. Dies bedeutet, dass der Basis-Score bei Null beginnt und durch jede valide Maßnahme leicht gesteigert werden kann. Die Auswahl dieses *Scoring Systems*, das in jeder Kategorie einen Wert von 0–5 vergibt, basiert auf der Annahme, dass Modelle, die nicht darauf trainiert wurden, robust zu handeln, nicht zufällig robust werden. Aktuelle Methoden, um KI-Modelle zu trainieren, wurden meist so entwickelt, dass sie die simpelsten, am stärksten ausgeprägten Muster in Datensätzen finden. Das sind meistens nicht die robustesten Muster. Hinzu kommt, dass für das Finden robuster Muster im Normalfall mehr Daten und eine längere Entwicklungsphase benötigt wird, um das Modell auf die Datengegebenheiten einzustellen.

Analog zu der Gewichtung der Fragen wurde auch den Antwortmöglichkeiten der Fragen jeweils ein Faktor zwischen -1 und +2 zugewiesen, der während des Assessments durch eines der folgenden Symbole dargestellt wird: -, o, +, ++. Dieser Faktor wird mit der Gewichtung der jeweiligen Frage multipliziert und fließt so gewichtet in das Gesamtergebnis ein. Für die Einschätzung der benötigten Robustheit erhalten Antworten, die auf eine hohe benötigte Robustheit hinweisen einen hohen Faktor. Bei der Einschätzung der umgesetzten Maßnahmen erhalten Antworten, die auf die Umsetzung valider Maßnahmen hindeuten ebenfalls einen hohen Faktor.

Um die beiden Ergebniswerte – den *Required Robustness Score* und den *Actual Robustness Score* – zwischen 0 und 5 zu erhalten, werden alle Fragenergebnisse unter der jeweiligen Kategorie aufsummiert.

EVALUATION EXAMPLE



11 questions for required level and 24 questions for actual level were answered

$$\textit{Single Question Score} = \textit{Question weight} * \textit{Answer Value}$$

$$\textit{Required Robustness Score} = \sum \textit{Single Question Score}$$

4

Content of the Assessment

Inhalt des Assessments

4.1. REQUIRED ROBUSTNESS

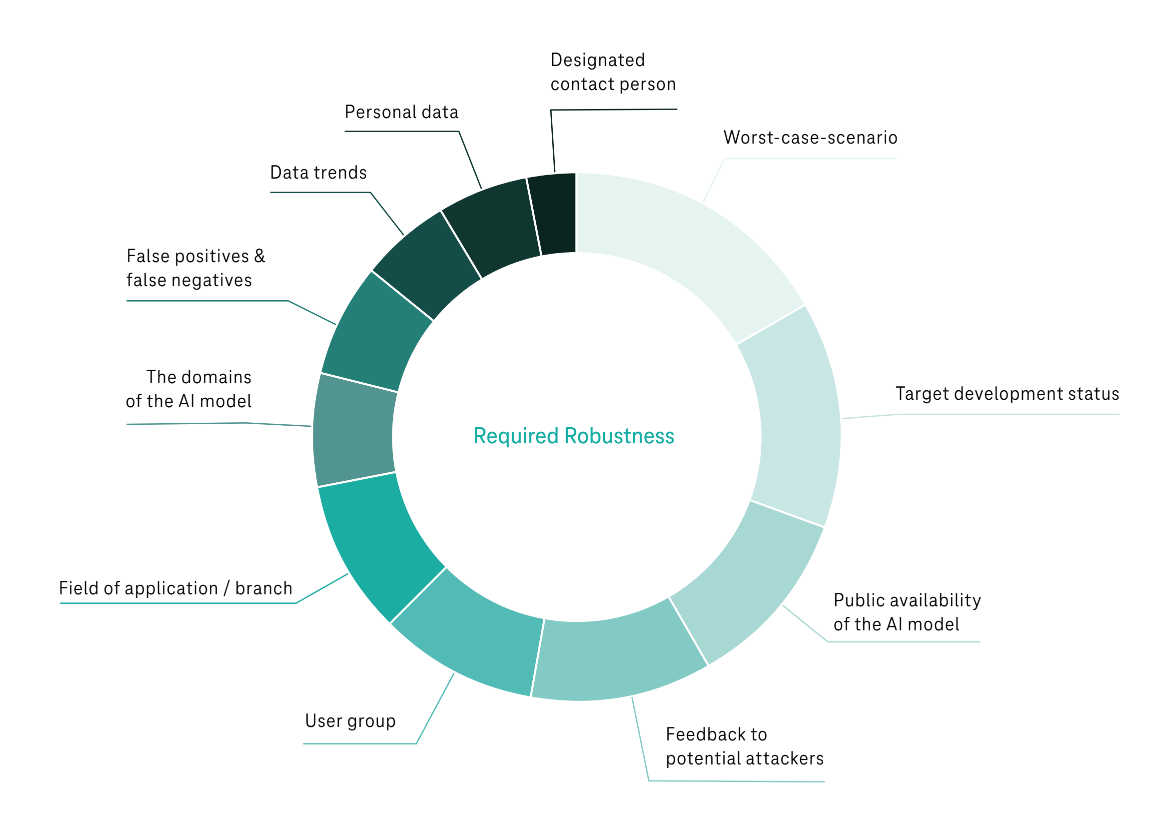
The goal of the first part of the assessment is to estimate how robust a solution should be for the given application. For this purpose, we used an approach based on the procedure of a risk analysis. Here, factors that estimate the effects of wrong decisions of the AI system and factors that show a probability of wrong decisions are considered. Both factors flow into *the Required Robustness Score* in approximately equal proportions through the selection and weighting of the questions. In the subsequent section, the weighting of the questions and the respective answer options are explained in detail.

A high value in the category "Required Robustness" indicates that the application requires a high degree of robustness.

4.1. REQUIRED ROBUSTNESS

Das Ziel des ersten Teilbereiches des Assessments ist die Einschätzung, wie robust eine Lösung für den gegebenen Anwendungsfall sein soll. Hierfür wird ein Ansatz verwendet, der an das Verfahren einer Risikoanalyse angelehnt ist. Betrachtet werden dafür Faktoren, die Auswirkungen von Fehlentscheidungen des KI-Systems einschätzen und Faktoren, die eine Wahrscheinlichkeit von Fehlentscheidungen aufzeigen. Beide Faktoren fließen über die Auswahl und Gewichtung der Fragen zu ungefähr gleichen Anteilen in den *Required Robustness Score* ein. Im Folgenden werden die Gewichtung der Fragen und die jeweiligen Antwortmöglichkeiten im Detail begründet.

Ein hoher Wert in der Kategorie *„Required Robustness"* lässt darauf schließen, dass der Anwendungsfall ein hohes Maß an Robustheit erfordert.



4.1.1. ASSESSING THE EFFECTS OF WRONG DECISIONS OF THE AI MODEL

4.1.1. EINSCHÄTZEN DER AUSWIRKUNGEN VON FEHLENTSCHEIDUNGEN DES KI-MODELLS

WORST-CASE-SCENARIO — 16.7 %

What kind of worst-case-scenario could theoretically be triggered by the model making a wrong decision?

Answer	Response Factor
High impact: The emergence of the risk forces the company to change its objectives or strategy in the short-term. Example: A failure leads to network instability on a large scale.	2
Medium impact: The emergence of the risk requires medium-term changes to the company's objectives or strategy. Example: The error leads to reputational damage, which receives attention from leading media sources.	1
Trivial impact: No effect on the company value. Example: Internal processes cannot take place at the usual speed.	0

In order to be able to assess how great the effects of a wrong decision are, it is necessary to consider the risk potential of a wrong decision in the worst case. With a question weighting of 16.7 %, the maximum hazard potential therefore has the greatest influence within the subcategories of possible effects as well as in the entire category of *Required Robustness*. The response options focus on the consequences for the company and provide an indicator for assessment. Being forced to change the company's goals or strategy in the short term, for example in the case of large-scale network instability, is a worst-case scenario with a high impact on the company. At the same time, a failure which receives attention from leading media sources is rated as having a medium impact on the company. Small mistakes that do not affect reputation, goals, and strategy are considered trivial.

WORST-CASE-SZENARIO — 16,7 %

Welche Art von Worst-Case-Szenario könnte theoretisch durch eine Fehlentscheidung des Modells ausgelöst werden?

Antwort	Antwort-Faktor
Hoher Einfluss: Das Auftreten des Risikos zwingt das Unternehmen, seine Ziele oder Strategie kurzfristig zu ändern. Beispiel: Ein Ausfall führt zu Netzinstabilität im großen Maß.	2
Mittlerer Einfluss: Das Auftreten des Risikos erfordert mittelfristige Änderungen der Ziele oder der Strategie des Unternehmens. Beispiel: Der Fehler führt zu einem Reputationsschaden, welcher von führenden Medien aufgegriffen wird.	1
Trivialer Einfluss: Keine Auswirkung auf den Unternehmenswert. Beispiel: Interne Prozesse können nicht in gewohnter Geschwindigkeit erfolgen.	0

Um einschätzen zu können, wie groß die Auswirkungen einer Fehlentscheidung sind, muss betrachtet werden, welches Gefahrenpotenzial bei einer Falsch-Entscheidung im schlimmsten Fall existiert. Das maximale Gefahrenpotenzial hat daher mit einer Fragengewichtung von 16,7 % den größten Einfluss innerhalb der Unterkategorien möglicher Auswirkungen sowie in der gesamten Kategorie *Required Robustness*. Die Antwortmöglichkeiten fokussieren die Folgen für das Unternehmen und geben einen Indikator zur Einschätzung. Der Zwang, kurzfristig die Unternehmensziele oder -strategie zu ändern, wie es zum Beispiel bei einer Netzinstabilität im großen Maß der Fall ist, ist ein Worst-Case-Szenario mit hohem Einfluss auf das Unternehmen, während ein Fehler, der von führenden Medien aufgegriffen wird, als mittelgroßer Einfluss auf das Unternehmen gewertet wird. Kleine Fehler, die sich nicht auf den Ruf, Ziele und Strategie auswirken, werden als trivial eingeschätzt.

▲ = Weighting of questions / Fragengewichtung

TARGET DEVELOPMENT STATUS — 13.9 %

What stage of development should the AI system reach?

Answer	Response Factor
Proof of concept	0
Finished product	1

Since the worst possible effects of a wrong decision rarely occur, it is also important to consider the effects of a more minor wrong decision. As the selection of potential consequences cannot be queried directly, due to the extreme variety of fields of application and possibilities, the possible fields of application are separated according to their respective development status in order to derive an estimate. AI systems that are developed as a *proof-of-concept* to demonstrate a certain functionality tend to have less impact in the case of a wrong decision, because they are often not connected to production systems but to a test environment. In addition, the results of a proof-of-concept are usually viewed in a more differentiated and critical way than the output of established and integrated systems, although these, too, can be susceptible.

The question was given a high weighting, as productive systems can cause significantly more major damage than *proof-of-concept* systems. However, the derivation of the risk is less direct than when the worst possible impact is queried and is therefore less clear-cut. For this reason, the question value was estimated to be somewhat lower at 13.9 %.

ZIEL-ENTWICKLUNGSSTAND — 13,9 %

Welchen Entwicklungsstand soll das KI-System erreichen?

Antwort	Antwort-Faktor
Proof of Concept	0
Fertiges Produkt	1

Da die schlimmsten möglichen Auswirkungen einer Fehlentscheidung nur sehr selten tatsächlich eintreten, sollte ebenfalls betrachtet werden, welche Auswirkungen eine kleinere Fehlentscheidung haben kann. Da durch die extreme Vielzahl an Einsatzbereichen und -möglichkeiten die Auswahl potenzieller Folgen kaum direkt abfragbar ist, werden stattdessen die möglichen Einsatzbereiche nach ihrem jeweiligen Entwicklungsstand getrennt, um dadurch eine Schätzung abzuleiten. KI-Systeme, die als *Proof of Concept* entwickelt werden, um eine bestimmte Funktionalität zu demonstrieren, tendieren zu geringeren Auswirkungen bei einer Fehlentscheidung, da sie oft nicht an Produktivsysteme, sondern an eine Testumgebung angebunden sind. Zudem werden die Ergebnisse eines *Proof of Concepts* meist differenzierter und kritischer betrachtet als die Ausgaben etablierter und integrierter Systeme, obwohl auch diese anfällig sein können.

Der Frage wurde eine hohe Gewichtung zugewiesen, da produktive Systeme deutlich höhere Schäden verursachen können als *Proof-of-Concept*-Systeme. Die Ableitung des Risikos erfolgt aber weniger direkt als bei der Abfrage der schlimmsten möglichen Auswirkung und ist so weniger eindeutig. Daher wurde der Fragenwert mit 13,9 % etwas geringer eingeschätzt.

FIELD OF APPLICATION / BRANCH — 9.7 %

In which industry sector will the AI model be used?

Answer	Response Factor
High-risk sector (e.g. automotive sector, manufacturing industry, financial sector)	1
Other	0

Previous user and expert interviews have shown that different industries tend to have different critical *use cases*. The fact that most of the interview participants rated the significance of the use cases as less significant than, for example, the *worst-case scenario* or the targeted development status was due to the fact that even in industries with a tendency towards high risk potential, there are applications that are not very critical. An exact derivation to the specific use case was considered more difficult and therefore less accurate.

The response options are differentiated between domains with high risk potential and other domains. Domains in which many use cases have a high risk potential have been empirically identified as including the automotive, manufacturing, and financial sectors. This is based both on the prevailing opinion from the expert interviews and on research into potential risks in the various sectors.

EINSATZBEREICH / BRANCHE — 9,7 %

In welchem Industriesektor soll das KI-Modell eingesetzt werden?

Antwort	Antwort-Faktor
Hoch-Risiko-Bereich (z. B. Automobilsektor, herstellende Industrie, Finanzsektor)	1
Andere	0

In vorhergehenden Interviews mit Expertinnen und Experten sowie Nutzerinnen und Nutzern hat sich herausgestellt, dass unterschiedliche Branchen zu unterschiedlich kritischen *Use Cases* tendieren. Dass die meisten Interviewten die Aussagekraft aber geringer einschätzten als zum Beispiel das *Worst-Case-Szenario* oder den angestrebten Entwicklungsstand, lag daran, dass auch in Branchen mit tendenziell hohem Gefahrenpotenzial Anwendungen existieren, die wenig kritisch sind. Eine genaue Ableitung auf den spezifischen *Use Case* wurde als schwieriger und daher weniger genau eingeschätzt.

Bei den Antwortmöglichkeiten wird zwischen Domänen mit hohem Gefahrenpotenzial und anderen Domänen unterschieden. Als Domänen, in denen viele Anwendungsfälle ein hohes Gefahrenpotenzial haben, wurden unter anderem die Automobilbranche, die verarbeitende Industrie und der Finanzsektor empirisch identifiziert. Dies stützt sich sowohl auf die vorherrschende Meinung aus den Interviews als auch auf eine Recherche von potenziellen Risiken in den verschiedenen Branchen.

PERSONAL DATA — 5.6 %

Are trade secrets or is personal or personally identifiable information processed by the AI model?

Answer	Response Factor
Yes	1
No	0

AI models are optimized in the training phase to map data-based functions. This can lead to the integration of information beyond abstractions and generalized patterns into the training. Especially in very complex models with many trainable parameters, it is often the case that certain information is implicitly stored in the model. In academic experiments it has already been shown that it is theoretically possible to extract this implicitly stored information through data extraction attacks.[15]

Such an attack would be particularly critical if the model were trained using personal or personally identifiable data that could be made public in this way. This must be considered, particularly in conjunction with the personal rights granted by the GDPR. Since the damage to reputation and financial losses would be significantly greater than with most other data in the event of such a violation of personal rights, the handling of personal data is explicitly included. As the processing of trade secrets is associated with a similarly high potential for damage, these are also taken into consideration.

PERSONENBEZOGENE DATEN — 5,6 %

Werden Geschäftsgeheimnisse, personenbezogene oder personenbeziehbare Daten von dem KI-Modell prozessiert?

Antwort	Antwort-Faktor
Ja	1
Nein	0

KI-Modelle werden in der Trainingsphase daraufhin optimiert, datenbasierte Funktionen abzubilden. Dabei kann es dazu kommen, dass nicht nur Abstraktionen und generalisierte Muster antrainiert werden. Gerade bei sehr komplexen Modellen mit vielen trainierbaren Parametern kommt es auch oft dazu, dass bestimmte Informationen implizit im Modell abgespeichert werden. In akademischen Versuchen wurde bereits gezeigt, dass es theoretisch möglich ist, diese implizit gespeicherten Informationen durch Data-Extraction-Angriffe zu extrahieren.[15]

Sehr kritisch wäre ein solcher Angriff, wenn das Modell mithilfe von personenbezogenen oder personenbeziehbaren Daten trainiert wurde, die so publik werden könnten. Dies ist besonders im Zusammenspiel mit den durch die DSGVO gewährten Persönlichkeitsrechten zu betrachten. Der Reputations- und finanzielle Schaden wäre bei solch einer Verletzung von Persönlichkeitsrechten deutlich größer als bei den meisten anderen Daten. Deshalb wird der Umgang mit personenbezogenen Daten explizit mit aufgenommen. Da mit dem Prozessieren von Geschäftsgeheimnissen ein ähnlich hohes Schadenspotenzial einhergeht, werden diese ebenfalls betrachtet.

15 Jagielski, Matthew, et al. "High-fidelity extraction of neural network models." arXiv preprint arXiv:1909.01838. 2019.

In this context, further questions were developed which can be used as a supplement, a completion, or a replacement. The selection of the appropriate question depends on the data to be processed. Especially for use cases with highly sensitive data, such as protected features related to age, disability/chronic illness, education, ethnicity, gender, religion/belief and sexual identity, further steps to increase robustness might be advisable.
However, this detail is not relevant to the Robust AI Assessment described here.

Are intimate personal data or customer details used for training?

- No
- Personal health data (medical records, etc.)
- Personal data with economic characteristics

Is training data used to enable conclusions about business models or business knowledge?

- No
- Data contains economic characteristics of customers and/or business transactions
- Data contains information on customer structures
- Data contains economic characteristics

Does the training data directly or indirectly contain particularly sensitive data ("protected features": age, disability/chronic illness, education, ethnic origin, gender, religion/belief, sexual identity)?

- Yes
- No

In diesem Zusammenhang wurden weitere Fragen entwickelt, welche als Ergänzung, zur Vervollständigung oder als Ersatz verwendet werden können. Die Auswahl der passenden Fragestellung ist abhängig davon, welche Daten prozessiert werden. Speziell bei Use Cases mit Verwendung von besonders sensiblen Daten, wie zum Beispiel die sogenannten „Protected Features", welche sich auf Alter, Behinderung oder chronische Erkrankung, Bildung, ethnische Herkunft, Geschlecht, Religion, Weltanschauung und sexuelle Identität beziehen und daher besonders zu schützen sind, könnten weitere Schritte zur Robustheitssteigerung ratsam sein. Diese Ausdetaillierung ist aber nicht für das hier beschriebene Robust AI Assessment relevant.

Werden intime Personen- oder Klardaten von Kundinnen und Kunden zum Training verwendet?

- Nein
- Personenbezogene Gesundheitsdaten (Krankenakten etc.)
- Personenbezogene Daten mit wirtschaftlichen Kenndaten

Werden Trainingsdaten verwendet, die Rückschlüsse auf Geschäftsmodelle oder Business-Wissen ermöglichen?

- Nein
- Daten enthalten wirtschaftliche Kenndaten von Kundinnen und Kunden und/oder Geschäftsvorfällen
- Daten enthalten Informationen über Strukturen von Kundinnen und Kunden
- Daten enthalten wirtschaftliche Kenndaten

Enthalten die Trainingsdaten direkt oder indirekt besonders schützenswerte Daten („Protected Features": Alter, Behinderung oder chronische Erkrankung, Bildung, ethnische Herkunft, Geschlecht, Religion, Weltanschauung, sexuelle Identität)?

- Ja
- Nein

Is the training data classified according to specific characteristics?

- No
- Based on specific characteristics (e.g. model of smartphone, number of people in the household)
- Specifically, on the basis of "protected features"

Should people who are equal in the decision-making process from a data point of view also necessarily receive equal decisions/classifications (e.g. for marketing purposes, when "protected features" must not be used and therefore the same advertising should be sent to customers with the same purchase history)?

- Yes, possible
- Yes, mandatory
- No, not intended
- No, must not happen under any circumstances

DESIGNATED CONTACT PERSON — 2.8 %

Is there a designated contact person who can be contacted in the case of a wrong decision?

Answer	Response Factor
No	1
Yes	0

The presence of a designated person in charge or contact person who can intervene in the case of a wrong decision also has a very small influence when estimating the amount of damage caused by wrong prediction. In these cases, a wrong decision cannot be prevented from the outset, but rather a contact person can be deployed to minimize damage, to mediate, and to clear up any misunderstandings as quickly as possible. The necessity of such a contact person has already been demonstrated in past projects and during preparatory expert interviews, but the contribution to the assessment of the general risk is rather small. Therefore, this variable has only been assigned a very low weighting.

Werden die Trainingsdaten anhand von speziellen Kenndaten in Klassen unterteilt?

- Nein
- Anhand von speziellen Eigenschaften (z. B. Modell des Smartphones, Anzahl der Personen im Haushalt)
- Gezielt anhand von „Protected Features"

Sollen Personen, die aus Datensicht dem Entscheidungsprozess gegenüber gleich erscheinen, auch zwingend gleiche Entscheidungen/Einordnungen erhalten (z. B. bei Marketingzwecken, wenn „Protected Features" nicht verwendet werden dürfen und daher Personen mit gleicher Einkaufshistorie die gleiche Werbung erhalten sollen)?

- Ja, möglich
- Ja, zwingend
- Nein, nicht so vorgesehen
- Nein, darf auf keinen Fall geschehen

DEDIZIERTE KONTAKTPERSON — 2,8 %

Gibt es eine dedizierte Kontaktperson, die bei einer falschen Entscheidung kontaktiert werden kann?

Antwort	Antwort-Faktor
Nein	1
Ja	0

Einen geringen Einfluss bei der Abschätzung der Schadenshöhe hat das Vorhandensein einer dedizierten Kontaktperson, die im Falle einer falschen Entscheidung eingreifen kann. Dadurch kann zwar die falsche Entscheidung nicht von vorneherein verhindert werden, jedoch kann solch eine Kontaktperson schadensminimierend eingesetzt werden, um zu vermitteln und um das Missverständnis möglichst schnell aufzuklären. Die Notwendigkeit einer Kontaktperson hat sich bereits in vergangenen Projekten und während der vorbereitenden Interviews bestätigt. Der Beitrag zur Einschätzung des allgemeinen Risikos ist hier allerdings eher gering. Deswegen wird nur eine sehr niedrigere Gewichtung vergeben.

4.1.2. ASSESSING THE PROBABILITY OF WRONG DECISIONS OF THE AI MODEL

PUBLIC AVAILABILITY OF THE AI MODEL — 11.1 %

Should the AI model and its underlying functionality be published?

Answer	Response Factor
Yes	1
No	0

This question aims to assess the extent to which an attacker is able to perform an *adversarial attack* to produce borderline cases. Since an attack using GAN *(Generative Adversarial Network)* is considered a very effective attack which often requires less effort from the attacker, the following section focuses on *adversarial attacks* using GAN.

To produce such an *adversarial attack*, the attacker develops their own AI model that produces data that is difficult for the attacked model to understand. First, the attacker model generates random data that is tested with the attacked model to see if the attacked model interprets the data correctly. Then the attacker model is iteratively optimized by machine learning methods such as gradient descent to generate data that are increasingly difficult to interpret.

If the AI model is publicly available, this gives the attacker the opportunity to replicate the model and prepare their attack, therefore managing to avoid detection before launching the attack. Since a targeted *adversarial attack* has the highest

4.1.2. EINSCHÄTZEN DER EINTRITTS-WAHRSCHEINLICHKEIT VON FEHL-ENTSCHEIDUNGEN DES KI-MODELLS

ÖFFENTLICHE VERFÜGBARKEIT DES KI-MODELLS — 11,1 %

Soll das KI-Modell und die zugrundeliegende Funktionsweise veröffentlicht werden?

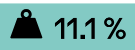
Antwort	Antwort-Faktor
Ja	1
Nein	0

Diese Frage zielt darauf ab, inwieweit es Angreifenden möglich ist, einen *Adversarial Attack* durchzuführen, um gezielt Grenzfälle zu produzieren. Da der Angriff mittels GAN *(Generative Adversarial Network)* als sehr effektiv gilt und zudem oft einen eher geringeren Aufwand für Angreifende darstellt, wird im Folgenden der Fokus darauf gelegt.

Um einen solchen *Adversarial Attack* zu produzieren, entwickeln Angreifende ein eigenes KI-Modell, das Daten produziert, die schwer für das angegriffene Modell zu verstehen sind. Zuerst generiert das angreifende Modell zufällige Daten, die mit dem angegriffenen Modell erprobt werden, um zu erkennen, ob dieses die Daten richtig interpretiert. Anschließend wird das angreifende Modell mittels maschinellem Lernverfahren – wie Gradient-Descent – iterativ daraufhin optimiert, Daten zu erzeugen, die immer schwieriger interpretierbar sind.

Wenn das KI-Modell öffentlich verfügbar ist, gibt dies den Angreifenden die Möglichkeit, das Modell zu replizieren und seinen Angriff vorzubereiten, ohne dass dies erkannt werden kann und bevor der tatsächliche Angriff stattfindet. Da ein zielgerichteter *Adversarial Attack* die

probability of causing wrong decisions and the publication of the model strongly favors this, this question has a relatively high weighting of 11.1 %. In the answer, a distinction can be made between a published and an internal/unpublished model.

FEEDBACK TO POTENTIAL ATTACKERS  11.1 %

Does a potential attacker receive feedback about which input results in which decision (output value)?

Answer	Response Factor
Yes	1
No	0

However, replication of a fully published AI model is not the only way a potential attacker can perform an *adversarial attack*. Instead, (1) during the attacker model's training, the model could be used directly on the attacked model to evaluate the results, or (2) a data set of tested input and output pairs could be created to train a third model and derive the behavior of the attacked model from the data base *(reverse engineering)*. Subsequently, the same process as with a replicated model is used. This way, even without direct access to the model to be attacked, it is possible to produce inputs that are difficult to understand and have a comparable probability to trigger borderline cases as described in the previous question, the high weighting of 11.1 % percent must therefore also be applied here.

The only way to completely rule out the possibility of an *adversarial attack* is to ensure that potential attackers cannot understand which input requires which decision. This means that the attacker will be

höchste Wahrscheinlichkeit hat, falsche Entscheidungen hervorzurufen und die Veröffentlichung des Modells dies stark begünstigt, hat diese Frage eine verhältnismäßig hohe Gewichtung von 11,1 %. In der Antwort kann zwischen einem veröffentlichten und einem internen / nicht veröffentlichten Modell unterschieden werden.

RÜCKMELDUNG AN POTENZIELLE ANGREIFENDE — 11,1 %

Erhalten Angreifende eine Rückmeldung darüber, welche Eingabe welche Entscheidung (Rückgabewert) nach sich zieht?

Antwort	Antwort-Faktor
Ja	1
Nein	0

Die Replikation eines vollständig publizierten KI-Modells ist jedoch nicht die einzige Möglichkeit, wie Angreifende einen *Adversarial Attack* durchführen könnten. Stattdessen könnte (1) während des Trainings des angreifenden Modells direkt das angegriffene Modell zur Bewertung der Ergebnisse genutzt werden, oder (2) es könnte ein Datensatz aus erprobten Ein- und Ausgabe-Paaren erstellt werden, um so ein drittes Modell zu trainieren und das Verhalten des angegriffenen Modells aus der Datenbasis abzuleiten *(Reverse Engineering)*. Anschließend wird derselbe Prozess wie mit einem replizierten Modell verwendet. Da so auch ohne direkten Zugriff auf das anzugreifende Modell schwer verständliche Eingaben produziert werden können, die eine vergleichbare Wahrscheinlichkeit haben, Grenzfälle auszulösen wie bei der vorherigen Frage beschrieben, muss hier ebenfalls die hohe Gewichtung von 11,1 % zum Tragen kommen.

Die einzige Möglichkeit vollständig auszuschließen, dass ein *Adversarial Attack* durchgeführt werden kann, ist es sicherzustellen, dass Angreifende nicht nachvollziehen

unable receive any information about the decision-making process and cannot create a database from which this process could be derived. The answer options for this question are limited to "Yes" (feedback available) and "No" (no feedback available).

USER GROUP  9.7 %

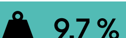
Who interacts with the AI model?

Answer	Response Factor
Unlimited user space (public)	2
Restricted user space (e.g. internal company)	1
Limited group of (trained) experts	0

Whether a model reaches an incorrect decision depends, to a large extent, on the user group interacting with the model. If the use case of the AI model requires that only a limited group of experts interact with the model, the probability of incorrect inputs and misinterpretations of outputs decreases. Nevertheless, the risk increases less than when information is disclosed through a public model or direct feedback loops, which is why the question weighting was placed lower than these two.

This question considers both the risk of natural variations in the input data and the risk of a targeted attack. On the one hand, the circle of potential attackers is automatically reduced for a small user group, which reduces the risk of an attack. On the other hand, the question aims at assessing whether a user has enough understanding to choose the right input and to critically question and interpret the result.

können, welche Eingabe welche Entscheidung bedingt. So erhalten Angreifende keine Information über den Entscheidungsfindungsprozess und können keine Datenbasis erstellen, aus der dieser Prozess abgeleitet werden könnte. Die Antwortmöglichkeiten beschränken sich bei dieser Frage auf „Ja" (Rückmeldung vorhanden) und „Nein" (keine Rückmeldung vorhanden).

NUTZUNGSGRUPPE  9,7 %

Wer interagiert mit dem KI-Modell?

Antwort	Antwort-Faktor
Uneingeschränkter Nutzungsraum (Öffentlichkeit)	2
Eingeschränkter Nutzungsraum (z. B. firmenintern)	1
Begrenzte Gruppe (geschulter) Expertinnen/Experten	0

Ob eine Fehlentscheidung eines Modells auftreten kann, hängt ebenfalls maßgeblich von der Nutzungsgruppe ab, die mit dem Modell interagiert. Wenn der Anwendungsfall des KI-Modells vorsieht, dass nur eine begrenzte Gruppe von Expertinnen und Experten mit dem Modell interagiert, sinkt die Wahrscheinlichkeit für fehlerhafte Eingaben und Fehlinterpretationen der Ausgaben. Trotzdem steigt das Risiko weniger als durch die Bekanntgabe von Informationen durch ein öffentliches Modell oder direkte Feedbackschleifen, weswegen die Fragengewichtung niedriger als diese beiden gewählt wurde.

Diese Frage berücksichtigt sowohl das Risiko von natürlichen Abweichungen in den Eingabedaten als auch das Risiko eines zielgerichteten Angriffes. Einerseits reduziert sich bei einer kleinen Nutzungsgruppe automatisch auch der Kreis der potenziellen Angreifenden, wodurch die Gefahr eines Angriffes sinkt. Andererseits zielt die Frage darauf ab, einzuschätzen, ob Nutzerinnen und Nutzer genug Verständnis haben, um die richtigen Eingaben zu wählen sowie das Ergebnis kritisch zu hinterfragen und zu interpretieren.

Only when users understand that AI models are stochastic models that usually draw the right conclusions and recognize real patterns, but will never be 100 percent error-free and adjust their expectations accordingly, will they be able to deal consistently with the potentially incorrect predictions. The differentiation of the response options divides the potential users into three groups with different expertise and training possibilities. While a group of in-house users can be informed and trained on certain specifics of the AI models, there is a much greater risk of inexperienced users in the unrestricted user space of the public. The least danger is posed by use by very limited expert groups, which can either be well prepared for a particular system or have expert knowledge in advance.

THE DOMAINS OF THE AI MODEL  6.9 %

Which domain does the AI model belong to?

Answer	Response Factor
High-risk domain (e.g. facial recognition, voice recognition and identification, predictive maintenance classifier/regression)	2
Medium-risk domain (e.g. image classifier, ASR for speaker/assistant, time series forecasting, reinforcement learning)	1
Low-risk domain (e.g. chatbot intent classifier, tabular data classifier/regression, simple classifier, simple multi-layer learning)	0

By using the domains of AI, two conclusions can be drawn regarding the consideration of risk. On the one hand, the type of model determines how strongly the model reacts to individual wrong decisions, and on the other hand, it allows us to estimate how complex or interpretable the input and output data will be.

Nur wenn Nutzerinnen und Nutzer verstehen, dass KI-Modelle stochastische Modelle sind, die zwar meist richtige Schlüsse ziehen und echte Muster erkennen, aber nie zu 100 % fehlerfrei sein werden und ihre Erwartungen dementsprechend anpassen, können sie konsequent mit möglicherweise falschen Vorhersagen umgehen. Die Unterscheidung der Antwortmöglichkeiten teilt die möglichen Nutzerinnen und Nutzer dabei in drei Gruppen mit unterschiedlicher Expertise und Schulungsmöglichkeiten. Während die firmeninterne Gruppe zu bestimmten Spezifika der KI-Modelle informiert und geschult werden kann, besteht im uneingeschränkten Nutzungsraum der Öffentlichkeit eine deutlich größere Gefahr unerfahrener Nutzerinnen und Nutzer. Die geringste Gefahr geht von einer Nutzung durch sehr limitierte Fachgruppen aus, die entweder gut auf ein bestimmtes System vorbereitet werden können oder schon im Vorhinein Fachwissen haben.



DIE DOMÄNE DES KI-MODELLS  6,9 %

Welcher Domäne ist das KI-Modell zugehörig?

Antwort	Antwort-Faktor
Hoch-Risiko-Domäne (z. B. Gesichtserkennung, Spracherkennung und -identifizierung, Predictive Maintenance Classifier/Regression)	2
Mittel-Risiko-Domäne (z. B. Bild-Klassifizierung, ASR für Speaker/Assistant, Zeitreihen-Vorhersage, Reinforcement Learning)	1
Niedrig-Risiko-Domäne (z. B. Chatbot-Absichtsklassifizierung, Tabellendatenklassifizierung/Regressionsaufgaben, einfacher Klassifizierer, einfaches mehrschichtiges Lernen)	0

Mithilfe der Domäne der KI, können für die Betrachtung des Risikos zwei Rückschlüsse gezogen werden. Einerseits leitet sich aus der Art des Modells ab, wie stark das Modell auf einzelne Fehlentscheidungen reagiert und andererseits kann man so abschätzen, wie komplex oder interpretierbar die Ein- und Ausgabedaten sein werden. Die Aussagekraft

However, the significance of these findings is limited by the fact that the task is only a rough indicator of the risk and the actual data is not directly included in this consideration. In order to do this, a much more nuanced and elaborate preliminary analysis would be necessary. Therefore a weighting of 6.9 % is assigned here.

Tasks such as clustering, which is used to divide a large amount of data into subgroups, are much less susceptible to individual data altered by an *adversarial attack* than tasks that are trained using data & labels (*supervised learning*). This is because a single input in clustering is assigned less information content than other types of tasks.

With the increasing number of attributes in highly complex tasks such as image or speech recognition, the probability of detecting a change in input data at an early stage decreases significantly. Thus, both natural changes or errors in the input data and specifically manipulated inputs tend to be detected later. This increases the probability of late detection of wrong decisions and increases the amount of time until countermeasures can be applied.
We identified three different classes for evaluating decisions:

1. **Low risk** for tasks that do not react strongly to individual changed data or have very complex input data, such as clustering or recommendation systems

2. **Medium risk** for tasks that either react strongly to individual changed data or have very complex input data, such as classic regression or classification tasks

3. **High risk** for tasks that react strongly to single changed data and have very complex input data, such as image recognition or speech recognition

dieser Erkenntnisse ist allerdings dadurch begrenzt, dass die Aufgabe nur ein grober Indikator des Risikos ist und die tatsächlichen Daten nicht direkt in diese Betrachtung einbezogen werden. Um dies zu tun, wäre eine deutlich nuancierte, aufwändigere Vorbetrachtung notwendig. Daher wird hier eine Gewichtung von 6,9 % vergeben.

Aufgaben, wie zum Beispiel das Clustering, das genutzt wird, um eine große Anzahl an Daten in Untergruppen aufzuteilen, reagieren deutlich weniger anfällig auf einzelne, durch einen *Adversarial Attack* veränderte Daten, als Aufgaben, die anhand von Daten und Labeln *(Supervised Learning)* trainiert werden. Das liegt daran, dass einer einzelnen Eingabe im Clustering weniger Informationsgehalt zugewiesen wird als bei anderen Aufgabentypen.

Mit der steigenden Anzahl an Attributen bei hochkomplexen Aufgaben wie Bild- oder Spracherkennung sinkt die Wahrscheinlichkeit stark, dass eine Änderung der Eingabedaten frühzeitig bemerkt wird. So werden sowohl natürliche Änderungen oder Fehler in den Eingabedaten als auch gezielt manipulierte Eingaben tendenziell später erkannt. Dadurch steigt die Wahrscheinlichkeit, dass Fehlentscheidungen erst später erkannt werden und es lange dauert bis gegengesteuert werden kann.
Daraus leiten sich drei verschiedene Klassen ab, die für die Entscheidung genutzt werden:

1. **Niedriges Risiko** für Aufgaben, die weder stark auf einzelne geänderte Daten reagieren noch sehr komplexe Eingabedaten haben, wie zum Beispiel Clustering oder Empfehlungssysteme

2. **Mittleres Risiko** für Aufgaben, die entweder stark auf einzelne geänderte Daten reagieren oder sehr komplexe Eingabedaten haben, wie zum Beispiel klassische Regressions- oder Klassifikationsaufgaben

3. **Hohes Risiko** für Aufgaben, die sowohl stark auf einzelne geänderte Daten reagieren als auch sehr komplexe Eingabedaten haben, wie zum Beispiel Bilderkennung oder Spracherkennung

FALSE POSITIVES & FALSE NEGATIVES


6.9 %

Are there strong differences in meaning between false negative and false positive decisions?

Answer	Response Factor
Yes	1
No	2

The expert interviews conducted during the preliminary review have shown that in many high-risk use cases there are strong differences between the significance of false-positive and false-negative results. A widely known example of this is the decision of whether to administer a (virtually side-effect free) drug for a severe disease. In this case, the decision not to administer a drug would be significantly worse than to administer a drug incorrectly even though no disease is present. The difference in meaning between a false-positive decision (drug without disease) and a false-negative decision (no drug despite disease) is therefore extremely significant.

FALSCH-POSITIVE FALSCH-NEGATIVE ENTSCHEIDUNGEN

6,9 %

Gibt es starke Bedeutungsunterschiede zwischen falsch-negativen und falsch-positiven Entscheidungen?

Antwort	Antwort-Faktor
Ja	1
Nein	2

Innerhalb der Interviews mit den Expertinnen und Experten während der Vorbetrachtung hat sich gezeigt, dass in vielen Anwendungsfällen mit hohem Risiko starke Unterschiede zwischen der Bedeutung von falsch-positiven und falsch-negativen Ergebnissen auftreten. Ein weitreichend bekanntes Beispiel dafür ist die Entscheidung, ob ein (nahezu nebenwirkungsfreies) Medikament gegen eine starke Krankheit verabreicht werden soll. Dabei wäre die Entscheidung kein Medikament zu verabreichen deutlich schlimmer, als fälschlicherweise ein Medikament zu verabreichen, obwohl keine Krankheit vorliegt. Der Bedeutungsunterschied einer falsch-positiven Entscheidung (Medikament ohne Krankheit) und einer falsch-negativen Entscheidung (kein Medikament trotz Krankheit) ist also groß.

A Confusion Matrix illustrates the effects of all possible (wrong) decisions of a model. Only when the value of a certain decision has been determined can it be judged whether the decisions are good enough for the respective application.

Eine Confusion-Matrix veranschaulicht die Auswirkungen aller möglichen (Fehl-)Entscheidungen eines Modells. Erst wenn festgestellt wird, welchen Wert eine bestimmte Entscheidung hat, kann beurteilt werden, ob die Entscheidungen gut genug für den jeweiligen Anwendungsfall sind.

Type of Prediction

	Positive Prediction	Negative Prediction
Positive Target	True-Positive	False-Negative
Negative Target	False-Positive	True-Negative

Value of Prediction

	Positive Prediction	Negative Prediction
Positive Target	+5 Appropriate medicine can be given	-7 Serious illness without medication
Negative Target	-2 Unnecesary medication	+2 End of quarantine

In addition to this identified trend, such a gradient makes it more difficult to assess the decision quality of the underlying model. An accuracy assessment does not include the type of error. Thus, serious errors (no medication despite illness) and less serious errors (unnecessary medication) are treated equally. This distorts the measurement.

Zusätzlich zu diesem identifizierten Trend macht ein solches Gefälle die Bewertung der Entscheidungsgüte des zugrundeliegenden Modells schwieriger. Bei einer Bewertung mittels Akkuranz wird nicht die Art des Fehlers einbezogen. So werden gravierende Fehler (kein Medikament trotz Krankheit) und weniger gravierende Fehler (unnötiges Medikament) gleichbehandelt. Das verzerrt die Messung.

A seemingly very high accuracy of, for example, over 90 % correct decisions, should not be seen as an indication of a well-functioning model. The reason for this can again be shown by the example of a disease. A disease with which 1 % of the population is infected would be correctly identified by an AI model that always produces only the result "negative" in 99 % of cases – completely without decision logic. However, this information has no added value for the evaluation of the prediction quality.

Eine sehr hoch scheinende Akkuranz von beispielsweise weit über 90 % richtiger Entscheidungen darf nicht als Indiz für ein gut funktionierendes Modell gesehen werden. Der Grund dafür lässt sich wieder am Beispiel einer Krankheit aufzeigen. Eine Krankheit, mit der 1 % der Population infiziert ist, würde ein KI-Modell, das immer nur das Ergebnis „Negativ“ produziert, in 99 % der Fälle richtig identifizieren – komplett ohne Entscheidungslogik. Diese Information hat jedoch keinen Mehrwert für die Bewertung der Vorhersagequalität.

In the evaluation of whether the prediction quality of the model is sufficient for the respective application, the possible (wrong) decisions have to be taken: true-positive, true-negative, false-positive, false-negative – weighted according to the business value associated with each one.

In die Bewertung, ob die Vorhersagequalität des Modells ausreichend für den jeweiligen Einsatzzweck ist, müssen die möglichen (Fehl-) Entscheidungen miteinbezogen werden: wahr-positiv, wahr-negativ, falsch-positiv, falsch-negativ – gewichtet nach dem jeweils damit verbundenen Business-Wert.

The assessment of whether the predictions are accurate enough becomes much more complex and is therefore more prone to false evaluations. If these more comprehensive methods are not used, the risk of overestimating the accuracy of the AI model increases.

Die Einschätzung, ob die Vorhersagen genau genug sind, wird so deutlich komplexer und ist somit anfälliger für falsche Bewertungen. Wenn auf solche detaillierteren Methoden verzichtet wird, steigt die Gefahr, die Genauigkeit des KI-Modells zu überschätzen.

Since the trends identified in the interviews are not academically confirmed, the question of the actual significance of the context remains open. However, because there is a broad consensus that an increasing significance gradient of the various wrong decisions leads to a more difficult assessment, this point is nevertheless assigned a reduced weighting of 6.9 %.

Da die in den Interviews festgestellten Tendenzen nicht akademisch bestätigt sind, bleibt die Frage der tatsächlichen Aussagekraft des Zusammenhangs offen. Weil es aber weitgehenden Konsens darüber gibt, dass ein steigendes Bedeutungsgefälle der verschiedenen Fehlentscheidungen zu einer schwierigeren Beurteilung führt, wird dieser Punkt trotzdem mit einer reduzierten Gewichtung von 6,9 % aufgenommen.

DATA TRENDS  5.6 %

Are the input/output data subject to trends?

Answer	Response Factor
No systematic changes in input data over time	0
Some trends and systematic deviations can occur in the input data	1
Rapid systematic changes in the input data are possible	2

Use cases in which the data is strongly influenced by trends, i.e. independent systematic changes such as weather, seasonal user behavior (e.g. holiday shopping), or other phenomena, have a higher risk that the patterns learned at the beginning can no longer be used to make reliable decisions. Indicators in such a case would be shifts in the status of the data at specific points in time as a result of external influences.

Although AI models are trained during the learning process to learn only generalized patterns, the training data can quickly change towards not reflecting all laws of the real world, especially with more complex tasks with limited data availability. Since trends are often not immediately recognized as such and training data is often limited to a narrow time frame, it is possible that important information is not included in the training data.

Especially with cyclical trends, such as the influence of the seasons on the images from outdoor cameras, there is also the challenge of correctly diagnosing incorrect decisions, as errors occur periodically and then appear to be corrected. Conversely with continuous trends, a "threshold" is more likely to be exceeded after which the quality of the decision is noticeably reduced, which is usually easier to interpret. However, since the value ranges change

DATENTRENDS 5,6 %

Unterliegen die Eingabe-/Ausgabe-Daten Trends?

Antwort	Antwort-Faktor
Keine systematischen Änderungen der Inputdaten im Laufe der Zeit	0
Einige Trends und systematische Abweichungen können in den Eingabedaten auftreten	1
Rapide systematische Veränderungen in den Inputdaten sind möglich	2

Anwendungsfälle, bei denen die Daten stark durch Trends – also durch unabhängige systematische Änderungen wie Wetter, saisonales Nutzungverhalten (zum Beispiel „Weihnachtseinkäufe“) oder andere Phänomene – beeinflusst werden, haben ein höheres Risiko, dass die zu Beginn gelernten Muster nicht mehr genutzt werden können, um zuverlässige Entscheidungen zu treffen. Symptomatisch ist bei so einem Fall, dass sich die Datenlage abhängig von der Zeit durch externe Einflüsse verschiebt.

Obwohl KI-Modelle während des Lernprozesses daraufhin trainiert werden, möglichst generalisierte Muster zu lernen, kann es gerade bei komplexeren Aufgaben mit begrenzter Datenverfügbarkeit schnell dazu kommen, dass die Trainingsdaten nicht alle Gesetzmäßigkeiten der realen Welt abbilden. Da Trends oft nicht sofort als solche erkannt werden und daher Trainingsdaten oft nur auf einen engen Zeitraum begrenzt sind, kann es sein, dass wichtige Informationen nicht in den Trainingsdaten enthalten sind.

Besonders bei zyklischen Trends, wie der Einfluss der Jahreszeiten auf die Bilder von Außenkameras, besteht darüber hinaus die Herausforderung, fehlerhafte Entscheidungen richtig zu diagnostizieren, da Fehler periodisch auftreten und dann wieder behoben scheinen. Bei fortlaufenden Trends dagegen wird eher eine „Schwelle“ überschritten nach der die Entscheidungsgüte merkbar sinkt, was meist einfacher zu interpretieren ist. Da sich aber bei

significantly for both types of trends, there is generally the same risk of generating data that produces wrong decisions due to natural changes. The level of risk depends on the speed and extent to which the trends vary.

This question was assigned a low weighting of 5.6 %, since trend-affected data do not immediately lead to wrong decisions, but only increase the probability of a wrong decision in the long run. Especially with image recognition systems, trends such as day and night as well as summer and winter can obviously be diagnosed, which means that countermeasures can often be taken before a wrong decision is made.

beiden Trendarten die Wertebereiche signifikant ändern, besteht generell das gleiche Risiko, Daten zu generieren, die aufgrund von natürlichen Veränderungen Fehlentscheidungen produzieren. Die Höhe des Risikos hängt von Änderungsgeschwindigkeit und -ausmaß der Trends ab.

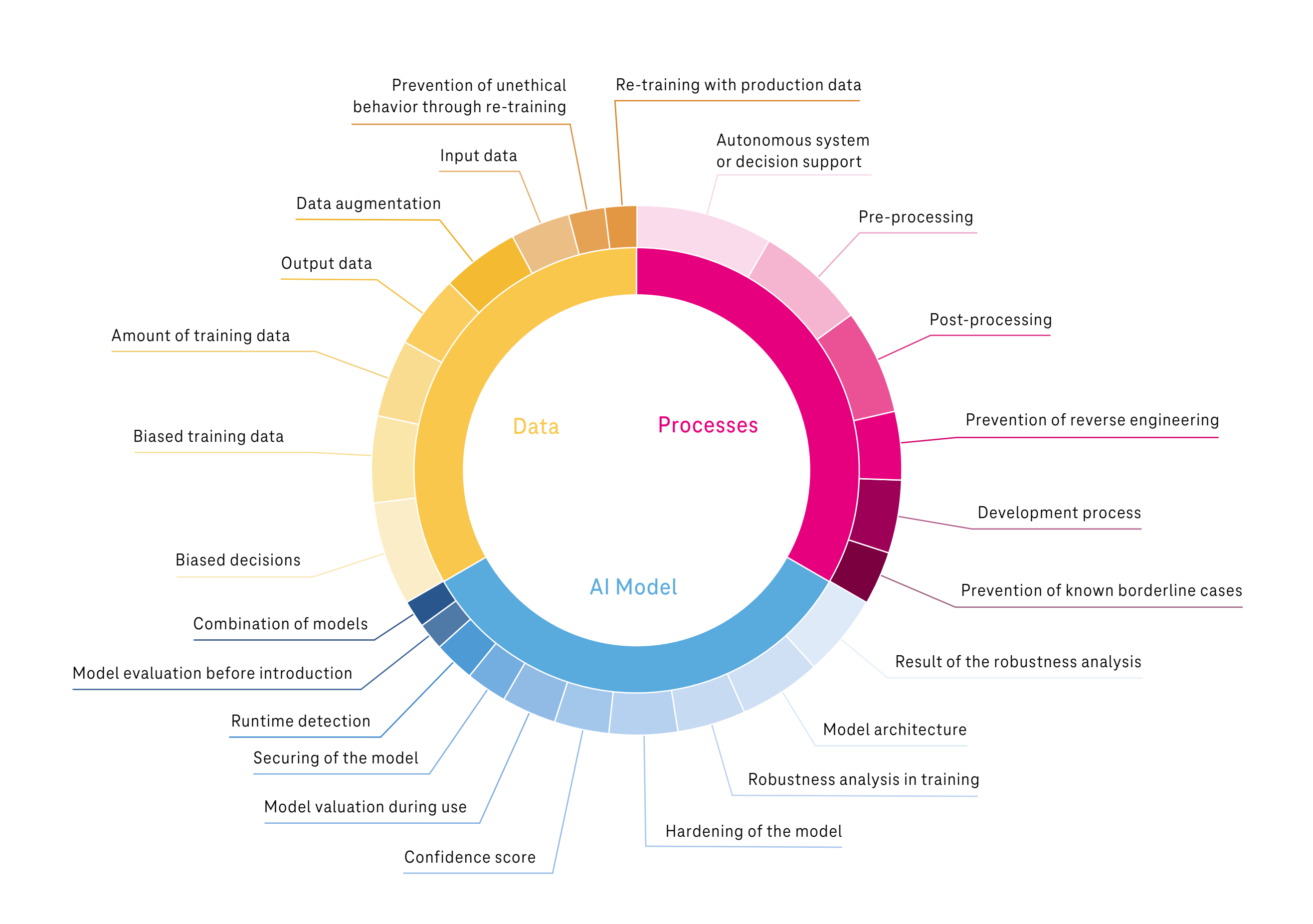
Diese Frage wurde mit einer niedrigen Gewichtung von 5,6 % aufgenommen, da trendbehaftete Daten nicht sofort zu Fehlentscheidungen führen, sondern lediglich die Wahrscheinlichkeit einer Fehlentscheidung langfristig wächst. Gerade bei Bilderkennungssystemen sind zudem Trends wie Tag und Nacht sowie Sommer und Winter offensichtlich zu diagnostizieren, wodurch oft schon gegengesteuert wird, bevor eine Fehlentscheidung auftritt.

4.2. ACTUAL ROBUSTNESS

The goal of the second part of the assessment is to assess how robustly a solution is implemented. For this purpose, the previously defined domains of robust AI are considered: (1) robust process, (2) robust AI model and (3) robust data basis. One third of each of these domains are included in the evaluation and thus form the *Actual Robustness Score*, which lies between the values 0 and 5. The higher the result value of this sub-range, the more robust the system under consideration is and the more suitable it is for use in critical application areas.

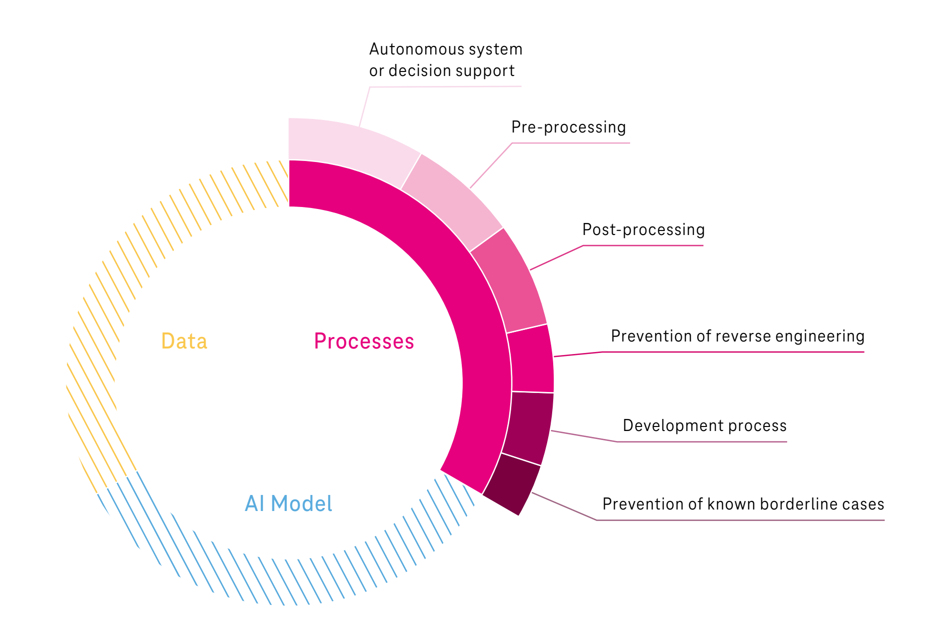
4.2. ACTUAL ROBUSTNESS

Das Ziel des zweiten Teilbereiches des Assessments ist die Einschätzung, wie robust eine Lösung tatsächlich umgesetzt ist. Betrachtet werden dafür die vorher definierten Domänen robuster KI: (1) robuster Prozess, (2) robustes KI-Modell und (3) robuste Datengrundlage. Die Teilbereiche fließen jeweils zu einem Drittel in die Bewertung ein und bilden damit den *Actual Robustness Score*, der zwischen den Werten 0 und 5 liegt. Je höher der Ergebniswert dieses Teilbereiches ausfällt, desto robuster ist das betrachtete System und desto besser ist es auch für den Einsatz in kritischen Anwendungsbereichen geeignet.



4.2.1. PROCESS

4.2.1. PROZESS



AUTONOMOUS SYSTEM OR DECISION SUPPORT — 25.8 %

Does the AI model make autonomous decisions or is it used as decision support for human decisions?

Answer	Response Factor
Decision support	1
Autonomous system	0

AUTONOMES SYSTEM ODER ENTSCHEIDUNGSHILFE — 25,8 %

Trifft das KI-Modell autonome Entscheidungen oder wird es als Entscheidungshilfe für menschliche Entscheidungen genutzt?

Antwort	Antwort-Faktor
Entscheidungshilfe	1
Autonomes System	0

An important aspect for the assessment of the robustness of an AI system is the process in which the underlying "intelligent" model is integrated. Even if an AI model implements all current technical measures to increase robustness, it can still be incontrovertibly assumed that the model will make wrong decisions. This is due to the stochastic nature of current AI models and is described by the "No

Ein wichtiger Aspekt für die Einschätzung der Robustheit eines KI-Systems ist der Prozess, in den das zugrundeliegende „intelligente" Modell eingebunden ist. Auch wenn ein KI-Modell alle aktuellen technischen Maßnahmen zur Steigerung der Robustheit implementiert, muss trotzdem angenommen werden, dass das Modell auch Fehlentscheidungen treffen wird. Dies ist in der stochastischen Natur aktueller KI-Modelle begründet und wird durch das

Free Lunch Theorem".[16] This theorem describes that the amount of potential input is nearly unlimited, but only a finite subset of it is correctly understood by a model. Any change to a model leads to a change of the range of understandable inputs. Nonetheless, there are still unlimited possibilities for unintelligible inputs, which will lead to wrong decisions.

Since it can never be ruled out that the model could make a wrong decision, it must be ensured that the surrounding process is flexible. One of the most effective ways to do this is to use the output as a decision aid for a human decision. Thus, while it can still lead to human error, it has had much less impact in the past because people usually subconsciously check their decisions for plausibility.

„No-Free-Lunch-Theorem“ beschrieben.[16] Dieses Theorem beschreibt, dass die Menge potenzieller Eingaben quasi unbegrenzt ist, jedoch nur ein begrenzter Teilbereich davon auch korrekt durch ein Modell verstanden wird. Jede Änderung eines Modells führt also nur dazu, dass sich der Bereich der verständlichen Eingaben verändert. Nebenher existieren allerdings weiterhin quasi unbegrenzte Möglichkeiten für unverständliche Eingaben, die zu Fehlentscheidungen führen können.

Weil also nie auszuschließen ist, dass das Modell eine Fehlentscheidung treffen kann, muss dafür gesorgt sein, dass der umschließende Prozess fehlerverzeihend ist. Eine der effektivsten Möglichkeiten dafür ist es, die Ausgabe als Entscheidungshilfe für eine Entscheidung durch einen Menschen zu nutzen. So kann es zwar immer noch zu menschlichen Fehlentscheidungen kommen, diese hatten in der Vergangenheit aber deutlich geringere Auswirkungen, da Menschen meist unterbewusst ihre Entscheidungen auf Plausibilität prüfen.

PRE-PROCESSING  19.4 %

What process does input data pass before it is processed by the AI model?

Answer	Response Factor
Complex pre-processing measures, e.g. balancing	2
Basic measures, e.g. standardization of data, plausibility, cleaning	1
No pre-processing	0

PRE-PROCESSING 19,4 %

Welchen Prozess durchlaufen Eingabedaten, bevor sie durch das KI-Modell prozessiert werden?

Antwort	Antwort-Faktor
Komplexe Pre-Processing-Maßnahmen, z. B. Balancing	2
Grundlegende Maßnahmen, z. B. Normierung der Daten, Plausibilitäten, Cleaning	1
Kein Pre-Processing	0

Another approach to reduce wrong decisions is the explicit restriction of the input and output possibilities of the AI system. By limiting the inputs, the virtually unlimited amount of unintelligible

Ein weiterer Ansatz, um Fehlentscheidungen zu reduzieren, ist die explizite Einschränkung der Ein- & Ausgabemöglichkeiten des KI-Systems. Durch die Beschränkung der Eingaben kann die quasi unbegrenzte Menge an unverständlichen potenziellen Eingaben deutlich verkleinert werden. So

16 Ho, Yu-Chi, and David L. Pepyne. "Simple explanation of the no free lunch theorem and its implications." Journal of optimization theory and applications: pp. 549–570. 15.3.2002.

potential inputs can be reduced significantly. This also reduces the risk that such inputs occur and lead to wrong decisions. This measure is mainly related to natural perturbation *(natural perturbation)* and has a negligible effect on the risk of an *adversarial attack*. The weighting was therefore set at 19.4 %.

Especially when it is easy to predict in which range of values valid inputs will likely move, a rigid restriction can be used to ensure that the inputs correspond with expectations and training. This increases the robustness considerably. If stochastic approaches are used for the pre-selection of the inputs, the robustness is also increased, but only to a small extent.

sinkt auch das Risiko, dass solche Eingaben auftreten und zu Fehlentscheidungen führen. Diese Maßnahme bezieht sich hauptsächlich auf natürliche Schwankungen und Eingabefehler *(Natural Perturbation)* und hat einen vernachlässigbaren Einfluss auf das Risiko eines *Adversarial Attacks*. Die Gewichtung wurde deshalb auf 19,4 % festgelegt.

Gerade, wenn gut vorherzusehen ist, in welchem Wertebereich sich valide Eingaben voraussichtlich bewegen werden, kann über eine starre Einschränkung sichergestellt werden, dass die Eingaben den Erwartungen und dem Training entsprechen. Dadurch steigt die Robustheit erheblich. Wenn stochastische Ansätze für die Vorauswahl der Eingaben genutzt werden, wird die Robustheit ebenfalls gesteigert – jedoch nur in geringem Ausmaß.

POST-PROCESSING  19.4 %

What process does input data go through after it has been processed by the AI model?

Answer	Response Factor
Human plausibility check or rigid, rule-based post-processing	2
Stochastic post-processing	1
No standardized process	0

POST-PROCESSING  19,4 %

Welchen Prozess durchlaufen Eingabedaten, nachdem sie durch das KI-Modell prozessiert wurden?

Antwort	Antwort-Faktor
Menschliche Plausibilitätsprüfung oder starres, regelbasiertes Post-Processing	2
Stochastisches Post-Processing	1
Kein standardisierter Prozess	0

By limiting the outputs, especially in the case of very dynamic outputs, a type of indirect plausibility check can be performed. This strengthens the robustness with regard to natural influences as well as to targeted *adversarial attacks*. However, for applications that only have a very limited number of possible outputs from the outset (e.g. a classification), only a small added value is created. However, since the majority of AI use cases produce diverse and dynamic outputs, the influence of post-processing can be regarded as relatively high. Therefore, the weighting was estimated to be 19.4 %.

Durch die Einschränkung der Ausgaben – gerade bei sehr dynamischen Ausgaben – kann eine Art indirekte Plausibilitätsprüfung durchgeführt werden. So wird die Robustheit sowohl in Bezug auf natürliche Einflüsse als auch in Bezug auf zielgerichtete *Adversarial Attacks* gestärkt. Bei Anwendungen, die von vorneherein nur eine sehr begrenzte Anzahl an möglichen Ausgaben haben, wie zum Beispiel einer Klassifizierung, entsteht allerdings nur ein geringer Mehrwert. Da aber die Mehrzahl der KI-Anwendungsfälle vielfältige und dynamische Ausgaben produziert, kann der Einfluss von Post-Processing als relativ hoch angesehen werden. Daher wurde die Gewichtung auf 19,4 % eingeschätzt.

Similar to pre-processing, a rigid rule-based estimation provides more robustness than stochastic estimations, and the answers were weighted accordingly.

PREVENTION OF REVERSE ENGINEERING — 12.9 %

How is the risk of an attacker being able to reverse engineer through repeated interactions reduced?

Answer	Response Factor
No strategy	0
Random switching between several models or similar strategy	1
Slow feedback loops	2

In order to prevent *adversarial attacks*, it is necessary to keep the internal functioning of the AI model as confidential as possible. As already described in previous questions, not only should the model itself not be publicly available, but it should also be made difficult to reproduce its functionality by repeated queries *(reverse engineering)*.

One way to counteract this is to limit the frequency of requests. If the frequency is chosen correctly, the use of the AI model is not affected – but *reverse engineering* is no longer possible in a realistic time horizon. In order to avoid negative effects for the user, it would be conceivable to gradually increase the slowdown after a certain number of requests with high speed.

Another way to make *reverse engineering* more difficult is to increase the complexity of the AI model. This can be achieved, for example, by randomly switching between different models that have been

Ähnlich wie beim Pre-Processing bietet eine starre, regelbasierte Einschätzung eine bessere Robustheit als stochastische Einschätzungen, dementsprechend wurden die Antworten gewichtet.

VORBEUGEN VON REVERSE ENGINEERING — 12,9 %

Wie wird das Risiko gesenkt, dass Angreifende durch wiederholte Interaktionen Reverse Engineering betreiben können?

Antwort	Antwort-Faktor
Keine Strategie	0
Zufälliges Wechseln zwischen mehreren Modellen oder ähnliche Strategie	1
Langsame Feedbackschleifen	2

Um *Adversarial Attacks* nachhaltig verhindern zu können, ist es nötig, dass die interne Funktionsweise des KI-Modells möglichst geheim bleibt. Wie bereits bei vorherigen Fragen beschrieben, sollte nicht nur das Modell an sich nicht öffentlich verfügbar sein, sondern es sollte auch erschwert werden, die Funktionsweise über wiederholtes Abfragen nachzuvollziehen (Reverse Engineering).

Eine Möglichkeit dem entgegenzuwirken ist, die Frequenz der Anfragen einzuschränken. Bei einer richtig gewählten Frequenz wird die Nutzung des KI-Modells nicht beeinträchtigt – ein *Reverse Engineering* ist dadurch aber in einem realistischen Zeithorizont nicht mehr möglich. Damit Nutzerinnen und Nutzer keine negativen Effekte spüren, wäre es hier denkbar, nach einer gewissen Anzahl an Anfragen mit hoher Geschwindigkeit, die Verlangsamung nach und nach zu verschärfen.

Eine andere Möglichkeit *Reverse Engineering* zu erschweren, ist die Komplexität des KI-Modells zu erhöhen. Das kann beispielsweise erreicht werden, indem zufällig

trained for the same task. Overall, however, the prevention of *reverse engineering* only helps against active *adversarial attacks* and not against *natural perturbation* in the input data. This partially limits the significance, which is why a weighting of 12.9 % was chosen.

DEVELOPMENT PROCESS  12.9 %

Has an extensive exploration phase been considered in the development process?

Answer	Response Factor
No explicit exploration phase, direct adoption of an academic model	-1
Short exploration phase with minor adjustments	0
Extensive exploration phase with extensive testing	1

Since many academic developments in the field of AI are applied in companies after a very short time, it is often difficult to assess whether these new/novel models are actually suitable for use in new products. In most cases, there is a lack of application-oriented tests to prove that the academically proven advantages indeed contribute to a better model in practice. Two challenges often arise:

1. Academic results are often very specific and not directly transferable. Since the configuration depends on the data used, adjustments are necessary to adapt the academic models to the requirements of the data and the use case. These adaptations are often limited to a few parameters, which, for example, change the structure and size of the model. However, they can have far-reaching effects that destabilize the training of the model, so that reliable learning is hardly possible. Several approaches should therefore be tested and

zwischen verschiedenen Modellen gewechselt wird, die für die gleiche Aufgabe trainiert wurden.
Insgesamt hilft das Vorbeugen von *Reverse Engineering* allerdings nur gegen aktive *Adversarial Attacks* und nicht gegen *Natural Perturbation* in den Eingabedaten. Dies schränkt die Aussagekraft teilweise ein, weswegen eine Gewichtung von 12,9 % gewählt wurde.

ENTWICKLUNGSPROZESS  12,9 %

Wurde eine ausgiebige Explorationsphase im Entwicklungsprozess berücksichtigt?

Antwort	Antwort-Faktor
Keine explizite Explorationsphase, direktes Übernehmen eines akademischen Modells	-1
Kurze Explorationsphase mit kleineren Anpassungen	0
Ausgiebige Explorationsphase mit umfangreichen Tests	1

Da im Bereich KI viele akademische Entwicklungen nach sehr kurzer Zeit Anwendung in Unternehmen finden, ist es oft schwer einzuschätzen, ob diese neuartigen Modelle tatsächlich für den Einsatz in neuen Produkten geeignet sind. Meist fehlen anwendungsnahe Tests, die zeigen, ob die akademisch bewiesenen Vorteile auch tatsächlich in der Praxis zu einem besseren Modell beitragen. Dabei kommen häufig zwei Herausforderungen auf:

1. Akademische Resultate sind oft sehr spezifisch und können nicht direkt übernommen werden. Da die Konfiguration von den verwendeten Daten abhängig ist, sind Anpassungen nötig, um die akademischen Modelle auf die Anforderungen der Daten und des Anwendungsfalls abzustimmen. Diese Anpassungen beschränken sich zwar oft auf wenige Parameter, die beispielsweise den Aufbau und die Größe des Modells ändern. Sie können aber dennoch weitreichende Auswirkungen haben, die das Training des Modells destabilisieren, sodass kaum zuverlässig gelernt werden kann. Es sollten daher mehrere Ansätze getestet

compared to achieve a good fit between data and model.

2. Due to the short development cycles, longer-term effects and interactions are often not known. These should at best be tested with prototypical developments and tests in the exploration phase.

The more extensive and open-ended the exploration phase, the greater the likelihood of creating a robust AI model well-matched to the data for the particular application. The total length of the exploration phase depends on the complexity of the solution and the application scenario. However, a long exploration phase is still no guarantee for a good or robust model. The weighting of the question was set at 12.9 %.

PREVENTION OF KNOWN BORDERLINE CASES — 9.7 %

How is the occurrence of already known borderline cases prevented?

Answer	Response Factor
No standardized process	0
During post-processing	1
During pre-processing	2

By accurately testing the AI model before introduction, the model's developers will already know in which situations the model works better or worse. The situations that are known to pose a particular challenge should be treated separately. This can be done either during *pre-processing* or during *post-processing.* The advantage of doing this during *pre-processing* is, on the one hand, to ensure that the AI model is not overloaded, and on the other

und verglichen werden, um eine gute Übereinstimmung zwischen Daten und Modell zu erreichen.

2. Durch die kurzen Entwicklungszyklen sind längerfristige Auswirkungen und Wechselwirkungen oft nicht bekannt. Diese sollten bestenfalls mit prototypischen Entwicklungen und Tests in der Explorationsphase erprobt werden.

Je umfangreicher und ergebnisoffener die Explorationsphase ist, desto größer ist die Wahrscheinlichkeit, dass ein robustes KI-Modell entsteht, das gut auf die Daten für den jeweiligen Anwendungsfall abgestimmt ist. Die absolute Länge der Explorationsphase ist abhängig von der Lösungskomplexität und des Einsatzszenarios. Insgesamt ist eine lange Explorationsphase trotzdem kein Garant für ein gutes oder robustes Modell. Die Gewichtung der Frage wurde auf 12,9 % festgelegt.

VORBEUGEN BEKANNTER GRENZFÄLLE — 9,7 %

Wie wird das Auftreten bereits bekannter Grenzfälle verhindert?

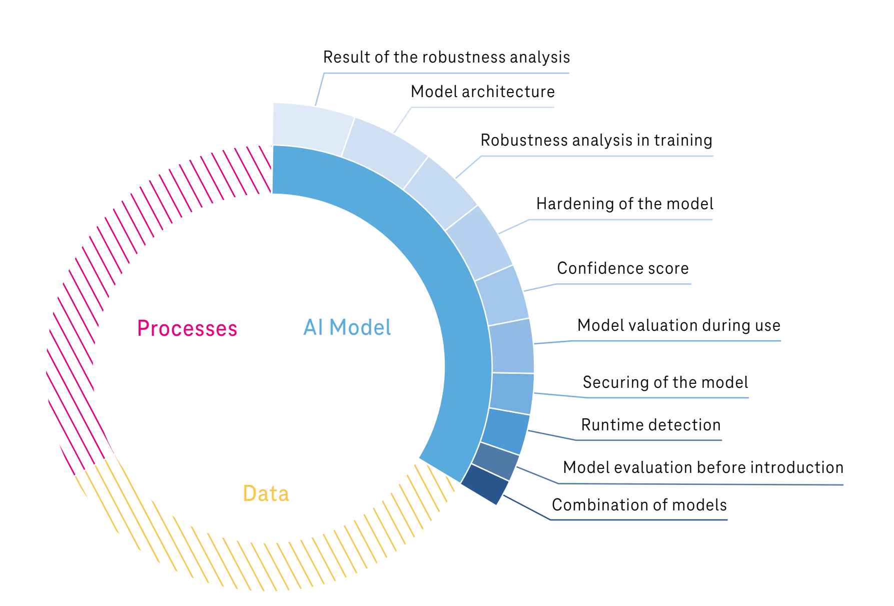
Antwort	Antwort-Faktor
Kein standardisierter Prozess	0
Während des Post-Processings	1
Während des Pre-Processings	2

Durch genaues Testen des KI-Modells vor der Einführung ist im besten Fall bereits bekannt, in welcher Situation das Modell besser funktioniert und in welcher schlechter. Die Situationen, die dafür bekannt sind, eine besonders große Herausforderung darzustellen, sollten separat behandelt werden. Dies kann entweder während des *Pre-Processings* oder während des *Post-Processings* durchgeführt werden. Der Vorteil dies während des *Pre-Processings* zu tun, liegt darin, dass einerseits das KI-Modell nicht beansprucht wird und andererseits, dass so die Eingabedaten – bevor

hand, to allow the input data to be modified before it is put into the AI model, in order to prevent the *edge case* and still allow processing.

Usually, however, only a limited number of actual edge cases are known in advance. In addition, it is often not trivial to reliably prevent these, so the weighting of the question is estimated to be lower than the other questions at 9.7 %.

4.2.2. AI MODEL



ROBUSTNESS ANALYSIS IN TRAINING — 12.5 %

Has the AI model been tested with unexpected inputs (such as an *adversarial attack*, noise, inputs with different context) during development?

Answer	Response Factor
No	0
Yes	1

sie in das KI-Modell gegeben werden – modifiziert werden können, damit der *Edge Case* verhindert wird und trotzdem eine Prozessierung stattfinden kann.

In der Regel ist jedoch nur eine begrenzte Menge der tatsächlichen *Edge Cases* vorher bekannt. Zusätzlich ist es oft nicht trivial diese zuverlässig zu verhindern, deswegen wird die Gewichtung der Frage mit 9,7 % niedriger eingeschätzt als die anderen Fragen.

4.2.2. KI-MODELL

ROBUSTHEITSANALYSE IM TRAINING — 12,5 %

Wurde das KI-Modell mit unerwarteten Eingaben (wie *Adversarial Attacks* , Rauschen, Eingaben mit anderem Kontext) während der Entwicklung getestet?

Antwort	Antwort-Faktor
Nein	0
Ja	1

The process of *adversarial attacks*, as described in the chapter Public availability of the AI model, can be used during training to analyze how the model is affected by such an attack. Results showing that the model is vulnerable to *adversarial attacks* are helpful for implementing necessary precautions based on this information. Focusing on a more robust process, rather than the derivation from such an analysis, contributes to a better robustness of the whole AI system. Since the susceptibility of the model to targeted attacks is an essential point, checking this is included with 12.5 % in the *actual robustness score.*

Der Prozess von *Adversarial Attacks*, wie er im Kapitel öffentliche Verfügbarkeit des KI-Modells beschrieben ist, kann während des Trainings dafür genutzt werden, um zu analysieren, wie das Modell durch einen solchen Angriff beeinflusst wird. Selbst Ergebnisse, die zeigen, dass das Modell anfällig für *Adversarial Attacks* ist, sind hilfreich, um basierend auf dieser Information andere Vorkehrungen zu treffen. Das Fokussieren auf einen robusteren Prozess als Ableitung aus solch einer Analyse trägt somit zu einer besseren Robustheit des gesamten KI-Systems bei. Da die Anfälligkeit des Modells gegenüber zielgerichteten Angriffen ein wesentlicher Punkt ist, fließt diese Überprüfung dessen mit 12,5 % in den *Actual Robustness Score* ein.

RESULT OF THE ROBUSTNESS ANALYSIS — 15.0 %

How does the AI model behave when it is deliberately misled by means of *adversarial attacks*?

Answer	Response Factor
Detected, system reacts as planned	1
Untargeted attacks can be successful	0
Targeted attacks can be successful	-1

ERGEBNIS DER ROBUSTHEITSANALYSE — 15,0 %

Wie verhält sich das KI-Modell, wenn es mittels *Adversarial Attacks* gezielt in die Irre geführt wird?

Antwort	Antwort-Faktor
Wird erkannt, System reagiert wie geplant	1
Nicht-zielgerichtete Angriffe können erfolgreich sein	0
Zielgerichtete Angriffe können erfolgreich sein	-1

Further points are awarded depending on the results of the robustness analysis. One indicator of robustness against adversarial attacks is the detection of targeted or untargeted attacks. Many recent studies have shown that unnoticed input changes can arbitrarily alter machine learning models' decision output.[17] These changes or misclassifications can be induced by targeted attacks and untargeted attacks, which differ in the following ways:

Je nachdem, wie die Robustheitsanalyse tatsächlich ausgefallen ist, werden weitere Punkte vergeben. Ein Indikator für die Robustheit gegen Adversarial Attacks ist das Erkennen von zielgerichteten, bzw. nicht-zielgerichteten Angriffen. Viele aktuelle Studien haben gezeigt, dass Entscheidungen, die von Machine-Learning-Modellen ausgegeben werden, durch unbemerkte Änderungen der Eingabe beliebig verändert werden können.[17] Diese Änderungen oder Misklassifizierungen können durch zielgerichtete oder nicht-zielgerichtete Angriffe herbeigeführt werden, wobei sich diese folgendermaßen unterscheiden:

17 Carlini & Wagner, 2017b; Szegedy et al., 2014

An untargeted attack is a source class misclassification attack that aims to misclassify the benign input by adding an interference. The input is then moved to a different, random class.

A targeted attack is also a misclassification, but this type of attack shifts the input to a specific class according to the attacker's intent.

Since it can be assumed that targeted attacks pursue a specific malicious goal with particular classifications, the weighting here is aligned towards this type of attack. Therefore, the result of the check influences the final result by 15.0 %.

Ein nicht-zielgerichteter Angriff ist ein Angriff zur Fehlklassifizierung der Ursprungsklasse, der darauf abzielt, die fehlerfreie Eingabe falsch zu klassifizieren, indem eine Störung hinzugefügt wird. Die Eingabe wird daraufhin in eine andere, willkürliche Klasse verschoben.

Bei einem zielgerichteten Angriff handelt es sich ebenfalls um eine Misklassifizierung. Allerdings wird bei einem solchen Angriff die Eingabe nach Absicht der Angreifenden in eine spezifische Klasse verschoben.

Da anzunehmen ist, dass zielgerichtete Angriffe ein bestimmtes bösartiges Ziel mit spezifischen Klassifizierungen verfolgen, wird hier die Gewichtung gezielt auf diese Art des Angriffs ausgerichtet. Daher beeinflusst das Ergebnis der Prüfung das Endergebnis mit 15,0 %.

MODEL ARCHITECTURE — 15.0 %

On which architecture is the applied AI model based?

Answer	Response Factor
Neural network	0
Conventional machine learning	1

MODELLARCHITEKTUR — 15,0 %

Auf welcher Architektur basiert das angewendete KI-Modell?

Antwort	Antwort-Faktor
Neuronales Netz	0
Herkömmliches Machine Learning	1

AI models are based either on conventional machine learning approaches or on a neural network. Conventional systems are significantly less complex and solve less complex tasks. On the one hand, the increased complexity of systems based on neural networks means that they can take on significantly more demanding tasks. On the other hand, this complexity makes it much more difficult to understand why certain decisions are made. Since the comprehension and verification of decisions is an essential part of robustness, this question is included in the evaluation with 15.0 %.

KI-Modelle basieren entweder auf herkömmlichen Machine-Learning-Ansätzen oder auf einem neuronalen Netz. Herkömmliche Systeme sind dabei deutlich weniger komplex und lösen weniger komplexe Aufgaben. Die gesteigerte Komplexität von Systemen, die auf neuronalen Netzen aufbauen, führt dazu, dass sie deutlich anspruchsvollere Aufgaben übernehmen können. Diese Komplexität macht es aber auch deutlich schwieriger nachzuvollziehen, weshalb bestimmte Entscheidungen getroffen werden. Da das Nachvollziehen und Verifizieren von Entscheidungen wesentliche Bestandteile der Robustheit sind, fließt diese Frage mit 15,0 % in die Bewertung ein.

HARDENING OF THE MODEL ▲ 12.5 %

Has the structure of the AI model been optimized for robustness or has the integrity of the existing data been evaluated (e.g. through several scaled image representations)?

Answer	Response Factor
Yes	1
No	0

HÄRTEN DES MODELLS ▲ 12,5 %

Wurde der Aufbau des KI-Modells für Robustheit optimiert oder die Integrität der vorhandenen Daten evaluiert (z. B. durch mehrere skalierte Bild-Repräsentationen)?

Antwort	Antwort-Faktor
Ja	1
Nein	0

Meanwhile, model architectures have been developed for various application areas to make models less sensitive to external influences and attacks. A well-known example is the use of multiple scaled inputs for the analysis of images. Because the model processes the image to be analyzed in multiple sizes, specific changes of individual pixels to confuse the model are less important. Since hardening of AI models is a current research area with many findings and changes, it is not possible to give a complete overview of possible measures here. As has been shown in many research projects, the application of hardening measures can reduce the risk of misjudgments.[18] The weighting for the use of hardening measures is included here with 12,5 %, because although a large potential theoretically exists, this is often not yet sustainably verified.

Für verschiedene Anwendungsbereiche wurden inzwischen Modellarchitekturen entwickelt, um Modelle unempfindlicher für äußere Einflüsse und Angriffe zu machen. Ein bekanntes Beispiel dafür ist das Nutzen von mehreren skalierten Eingaben für die Analyse von Bildern. Dadurch, dass das Modell das zu analysierende Bild in mehreren Größen prozessiert, fallen spezifische Änderungen einzelner Pixel, um das Modell zu verwirren, weniger stark ins Gewicht. Da das Härten von KI-Modellen ein aktuelles Forschungsgebiet mit vielen Erkenntnissen und Änderungen ist, kann hier kein gesamter Überblick über mögliche Maßnahmen gegeben werden. Wie bereits in der Forschung an vielen Stellen gezeigt wurde, kann die Anwendung von Härtungsmaßnahmen die Gefahr für Fehleinschätzungen reduzieren.[18] Die Gewichtung für die Nutzung von Härtungsmaßnahmen ist hier mit 12,5 % eingeflossen, da zwar theoretisch ein großes Potenzial vorhanden ist, dies aber oft noch nicht nachhaltig verifiziert.

18 Rosenfeld, Azriel, ed. al. Multiresolution image processing and analysis. Vol. 12. Springer Science & Business Media. 2013.

CONFIDENCE SCORE ▲ 10.0 %

Does the model produce a *Confidence Score* and how is it used?

Answer	Response Factor
No *Confidence Score*	0
For information purposes (e.g. for tracking)	1
Confidence Score has an influence on decision-making	2

CONFIDENCE SCORE ▲ 10,0 %

Produziert das Modell einen *Confidence Score* und wie wird dieser genutzt?

Antwort	Antwort-Faktor
Kein *Confidence Score*	0
Zu Informationszwecken (z. B. für die Nachverfolgung)	1
Confidence Score hat Einfluss auf Entscheidungsfindung	2

Although no direct probability can be deduced from a *Confidence Score* as to whether a particular model decision is right or wrong, the model can nevertheless use it to express its preference for or against a particular decision. Thus, one can ensure that the model makes decisions in which it is very confident. Even with high confidence values, wrong decisions cannot be ruled out. Here, however, they indicate deficits in the understanding of the model and thus provide an indicator of the type of inputs the model misunderstands. These findings can be used directly to improve training. Therefore, the *Robustness Score* also increases if the *Confidence Score* is "only" used as information and does not influence the model result. The existence and usefulness of a *Confidence Score* is included in the evaluation with 10.0 %.

Obwohl sich aus einem *Confidence Score* keine direkte Wahrscheinlichkeit ableiten lässt, ob eine bestimmte Modellentscheidung richtig oder falsch ist, kann dieser von dem Modell gleichwohl dazu genutzt werden, sich für oder gegen eine bestimmte Entscheidung auszusprechen. Man kann so also sicherstellen, dass das Modell Entscheidungen trifft, bei denen es sich sehr sicher ist. Auch bei hohen Confidence-Werten sind Fehlentscheidungen nicht ausgeschlossen. Hier deuten sie aber auf Defizite im Verständnis des Modells hin und geben so einen Indikator, welche Art Eingaben das Modell missversteht. Diese Erkenntnisse können direkt für eine Verbesserung des Trainings genutzt werden. Deshalb steigt der *Robustness Score* auch, wenn der *Confidence Score* „nur" als Information genutzt wird und nicht das Modellergebnis beeinflusst. Das Vorhandensein und Nutzen eines *Confidence Scores* fließt mit 10,0 % in die Bewertung ein.

MODEL VALUATION DURING USE ▲ 10.0 %

How is model quality monitored during model deployment?

Answer	Response Factor
No recurring quality inspection	0
Continuous monitoring of quality metrics or periodic assessments using sample data	1

MODELLBEWERTUNG WÄHREND NUTZUNG ▲ 10,0 %

Wie wird die Modellqualität während des Modelleinsatzes überwacht?

Antwort	Antwort-Faktor
Keine wiederkehrende Qualitätsprüfung	0
Durchgängiges Monitoring der Qualitätsmetriken oder periodische Assessments mittels Beispieldaten	1

As input data can change over time even without predictable trends, it has proven beneficial to examine how the quality of decisions changes over time. Whether this is done manually at regular intervals or by continuous assessment of automatic quality metrics was found to be irrelevant based on the previous expert interviews. If monitoring of this kind is done, it is possible to intervene at an early stage and prevent significant wrong decisions if a deterioration is detected. Therefore this question is included in the evaluation at 10.0 %.

Da sich auch ohne vorhersehbare Trends die Eingabedaten mit der Zeit verändern können, hat es sich als vorteilhaft erwiesen, zu prüfen, wie sich die Qualität der Entscheidungen mit der Zeit verändert. Ob dies in regelmäßigen Abständen händisch geprüft wird oder ob dies durch ein kontinuierliches Assessment automatischer Qualitätsmetriken geschieht, wurde von den interviewten Expertinnen und Experten als unerheblich befunden. Sofern eine Überwachung dieser Art geschieht, ist es möglich, früh einzugreifen und stärkere Fehlentscheidungen zu verhindern, wenn eine Verschlechterung festgestellt wird. Daher fließt diese Frage zu 10,0 % in die Bewertung ein.

SECURING OF THE MODEL 7.5 %

Is the model secured against espionage of model knowledge and training data?

Answer	Response Factor
No	0
Yes	1

"Poisoning" is used to specifically attack training data and contaminate it to retrain the model or disrupt retraining. By viewing the training data, the attacker can also derive conclusions about the model and its function, which could lead to further interference in its system. Securing the training data and model knowledge on the storage location with a security system such as a firewall influences the result by 7.5 %.

ABSICHERUNG DES MODELLS  7,5 %

Ist das Modell gegen Ausspähen von Modellwissen und Trainingsdaten gesichert?

Antwort	Antwort-Faktor
Nein	0
Ja	1

Durch das sogenannte „Poisoning" werden gezielt Trainingsdaten angegriffen und verunreinigt, um das Modell neu anzulernen oder das Neutraining zu stören. Außerdem können bei Einsicht in die Trainingsdaten Rückschlüsse auf das Modell und dessen Funktion getroffen werden, womit weiter in dessen System eingegriffen werden kann. Die Absicherung der Trainingsdaten und des Modellwissens auf dem Speicherort durch ein Sicherungssystem, wie zum Beispiel einer Firewall, beeinflusst das Ergebnis mit 7,5 %.

RUNTIME DETECTION  7.5 %

Does *Runtime Detection* check if the input is plausible?

Answer	Response Factor
No	0
Yes	1

Similar to the hardening of AI models, the detection of modified inputs is part of ongoing developments. Current promising approaches engage in the activation of neurons of neural networks during training and during specifically modified inputs. Even if the final result is the same for both, there are still differences in how the model arrives at the final decision. The model follows different paths to the same goal. By comparing these paths, an assessment can be made as to whether model inputs align with the training data. If there are major differences, the decision process can be aborted to prevent incorrect outputs.

Runtime Detection has proven to be helpful in academic tests, but few tests have been performed in practice. The weighting for this question is 7.5 % in the evaluation.

RUNTIME DETECTION  7,5 %

Wird mittels *Runtime Detection* geprüft, ob die Eingaben plausibel sind?



Antwort	Antwort-Faktor
Nein	0
Ja	1

Ähnlich wie das Härten von KI-Modellen ist auch die Erkennung von modifizierten Eingaben Teil von laufenden Entwicklungen. Aktuelle, vielversprechende Ansätze beschäftigen sich mit der Aktivierung der Neuronen von neuronalen Netzwerken während des Trainings und mit der Aktivierung der Neuronen bei gezielt modifizierten Eingaben. Auch wenn bei beiden das gleiche Endergebnis entsteht, gibt es dennoch Unterschiede, wie das Modell zu der Entscheidung am Ende kommt. Das Modell geht sozusagen unterschiedliche Pfade zum gleichen Ziel. Indem diese Pfade verglichen werden, kann eine Einschätzung abgegeben werden, ob Modelleingaben den Trainingsdaten ähnlich sind. Bei groben Abweichungen kann der Entscheidungsprozess abgebrochen werden, um Fehlausgaben zu verhindern.

Runtime Detection hat sich in akademischen Tests als hilfreich erwiesen, in der Praxis sind aber noch nicht viele Tests durchgeführt worden. Die Gewichtung fließt hier mit 7,5 % in die Bewertung ein.

MODEL EVALUATION BEFORE INTRODUCTION  5.0 %

How is the model validated before deployment?

Answer	Response Factor
Accuracy or comparable method	0
Comprehensive metrics such as AROC	1

MODELLBEWERTUNG VOR EINFÜHRUNG  5,0 %

Wie wird das Modell vor der Einführung validiert?

Antwort	Antwort-Faktor
Akkuranz oder vergleichbare Methode	0
Umfassende Metriken, wie z. B. AROC	1

In order to evaluate how well a prediction model performs, there are – as mentioned in the chapter *False-Positive & False-Negative* – different evaluation methods that can come to very different results when evaluating the same model.

Um zu bewerten, wie gut ein Vorhersagemodell ist, gibt es – wie im Kapitel *Falsch-Positive & Falsch-Negative* angeschnitten – verschiedene Bewertungsmethoden, die bei der Bewertung des gleichen Modells zu sehr unterschiedlichen Ergebnissen kommen können.

Accuracy = number of right decisions / number of wrong decisions

Akkuranz = Anzahl richtige Entscheidungen / Anzahl falsche Entscheidungen

As previously described, when only using accuracy as a means of evaluating model performance, it is not possible to recognize whether the evaluation result actually points to a good prediction model or whether a very advantageous database with a high base rate is available.

Wie bereits vorher beschrieben, ist es bei der Bewertung mittels Akkuranz nicht möglich zu erkennen, ob das Bewertungsergebnis tatsächlich auf ein gutes Vorhersagemodell hinweist oder ob eine sehr vorteilhafte Datenbasis mit hoher Base Rate vorliegt.

The base rate is the maximum accuracy a model can achieve if it always produces the same result. With two equally distributed decision options, this rate is 50 %. If the ratio is very unequal, the base rate is higher. If, in 90 % of the cases, decision 1 would be correct, a model can reach an accuracy of 90 % without needing any decision logic. Whether the accuracy of a model is good or bad always depends on the base rate.

Die Base Rate ist die maximale Akkuranz, die ein Modell erreichen kann, wenn es immer dasselbe Ergebnis produziert. Bei zwei gleichverteilten Entscheidungsmöglichkeiten liegt diese bei 50 %. Bei einem sehr ungleichen Verhältnis fällt die Base Rate höher aus. Wenn in 90 % der Fälle Entscheidung 1 korrekt wäre, kann ein Modell eine Akkuranz von 90 % erreichen, ohne jegliche Entscheidungslogik zu benötigen. Ob die Akkuranz eines Modells gut oder schlecht ist, hängt immer von der Base Rate ab.

Model evaluation functions such as the AROC (*Area under Receiver Operating Curve*) method, which refers to false positives and false negatives, allow better conclusions to be drawn about the actual prediction quality.

Modellbewertungsfunktionen, wie die AROC-Methode (Area under Receiver Operating Curve), nehmen Bezug auf Falsch-Positive und Falsch-Negative und lassen damit einen deutlich besseren Rückschluss auf die tatsächliche Vorhersagequalität zu.

AROC (Area under Receiver Operating Curve) is a model evaluation function often used in medical and pharmaceutical research. In this method, the sensitivity (true-positive rate) is compared with the specificity (false-positive rate). These are formed for different threshold values and plotted as a curve. The larger the area under this curve, the more meaningful the prediction model.

AROC (Area under Receiver Operating Curve) ist eine Modellbewertungsfunktion, die oft Verwendung in der Medizin- und Pharmaforschung findet. Bei dieser Methode wird die Sensitivität (Richtig-Positiv-Rate) mit der Spezifität (Falsch-Positiv-Rate) verglichen. Diese werden für verschiedene Schwellenwerte gebildet und als Kurve geplottet. Je größer die Fläche unter dieser Kurve ist, desto aussagekräftiger ist das Vorhersagemodell.

This question is included in the evaluation with a value of 5.0 %.

Diese Frage fließt mit einem Wert von 5,0 % in die Bewertung ein.

COMBINATION OF MODELS


5.0 %

Does the AI system consist of several linked models?

Answer	Response Factor
No	0
Yes, sequentially connected	1
Yes, connected in parallel *(Ensemble Learning)*	2

KOMBINATION AUS MODELLEN

5,0 %

Besteht das KI-System aus mehreren verknüpften Modellen?

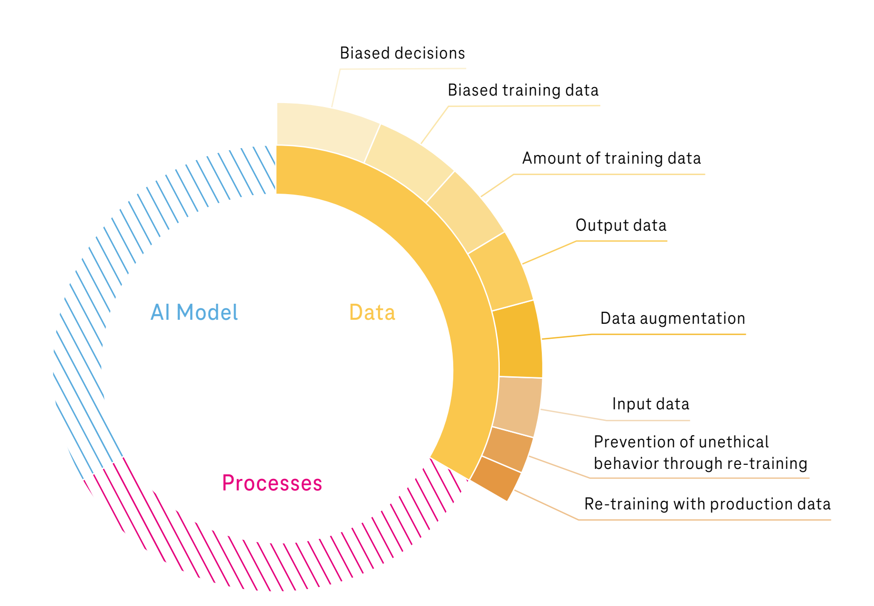
Antwort	Antwort-Faktor
Nein	0
Ja, sequentiell verbunden	1
Ja, parallel-geschaltet *(Ensemble Learning)*	2

A combination of several models helps to compensate for weaknesses of individual models. These models can be arranged either sequentially or in parallel. The advantage of parallel models is that they do not influence each other, so you can compare which outputs the models arranged next to each other have produced. If individual models produce different outputs than the rest of the models, it is easy to conclude that there is a high probability that wrong decisions were made. However, since all models are essentially based on the same existing data, they often have similar weaknesses. The situation where only individual models differ from a large majority of other models is therefore rare.
For this reason, the question is included in the evaluation with 5.0 %.

Eine Kombination mehrerer Modelle hilft dabei, Schwachstellen einzelner Modelle auszugleichen. Diese Modelle können entweder sequenziell oder parallel angeordnet werden. Parallel angeordnete Modelle haben den Vorteil, dass sie sich nicht gegenseitig beeinflussen und man so vergleichen kann, welche Ausgaben die nebeneinander angeordneten Modelle jeweils produziert haben. Wenn einzelne Modelle andere Ausgaben als die restlichen Modelle produzieren, kann man gut schlussfolgern, dass hier mit großer Wahrscheinlichkeit Fehlentscheidungen aufgetreten sind. Da alle dieser Modelle aber grundsätzlich auf den gleichen, vorhandenen Daten basieren, haben diese oft auch ähnliche Schwachstellen. Die Situation, dass sich nur einzelne Modelle von einer großen Mehrheit anderer Modelle unterscheiden, ist deshalb selten. Aus diesem Grund fließt diese Frage mit 5,0 % in die Wertung ein.

4.2.3. DATA

4.2.3. DATEN



BIASED DECISIONS	19.2 %

How do you prevent the AI model from making biased decisions?

Answer	Response Factor
No standardized process	0
Basic statistical analysis of the decisions	1
Sensitivity analyses of the model or special metrics to evaluate the decisions (e.g. SHAP value analysis, fairness metrics)	2

VOREINGENOMMENE ENTSCHEIDUNGEN	19,2 %

Wie wird verhindert, dass das KI-Modell voreingenommene Entscheidungen trifft?

Antwort	Antwort-Faktor
Kein standardisierter Prozess	0
Grundlegende statistische Analysen der Entscheidungen	1
Sensitivitätsanalysen des Modells oder spezielle Metriken zur Bewertung der Entscheidungen (z. B. SHAP-Value-Analyse, Fairness-Metriken)	2

AI models are often regarded as so-called "black box" models, because when a decision is made, it is not automatically clear which reasons have led to a particular decision. In order to ensure that model decisions are not based on bias or other ethically questionable motives, various analytical methods can be used. Normally, despite these analyses, part of the model's functionality remains unexplained, which is why we speak of *grey box* models (partially explainable) rather than *white box models* (fully explainable). The high weighting of this question (19.2 %) is due to the fact that these analyses are the only way to detect discrimination by AI with high probability. Measures that could be taken if the AI model has learned a certain bias include adjusting the training data or including fairness metrics in the model assessment method *(loss)* for re-training.

Depending on the type and complexity of the analysis, the proportion of the explainable behavior of the AI model changes. Statistical analyses can help to identify certain correlations. However, it is often the case that no direct causality can be deduced from these correlations. More complex approaches, such as SHAP value analysis, which uses an approach well-founded in game theory, can also provide conclusions about the causality of decisions.[19] Therefore, an additional category has been introduced here for the response options.

KI-Modelle werden oft als sogenannte *Black-Box-Modelle* angesehen, da bei einer Entscheidung nicht automatisch erkennbar ist, welche Gründe zu einer bestimmten Entscheidung geführt haben. Damit sichergestellt werden kann, dass die Modellentscheidungen nicht auf Voreingenommenheit oder auf anderen ethisch fragwürdigen Beweggründen basieren, können verschiedene Analysemethoden verwendet werden. Normalerweise bleibt trotz dieser Analysen ein Teil der Funktionsweise des Modells unerklärt, weswegen von *Grey-Box-Modellen* (teilweise erklärbar) und nicht von *White-Box-Modellen* (vollständig erklärbar) gesprochen wird. Die hohe Gewichtung dieser Frage (19,2 %) liegt daran, dass diese Analysen die einzige Möglichkeit sind, um Diskriminierung durch KI mit hoher Wahrscheinlichkeit zu erkennen. Maßnahmen, die ergriffen werden könnten, falls das KI-Modell eine bestimmte Voreingenommenheit gelernt hat, sind unter anderem das Anpassen der Trainingsdaten oder das Aufnehmen von Fairness-Metriken in die Modellbewertungsmethode *(Loss)* für ein erneutes Training.

Je nach Art und Komplexität der Analyse ändert sich der Anteil des erklärbaren Verhaltens des KI-Modells. Statistische Analysen können dazu beitragen, bestimmte Korrelationen erkennbar zu machen. Daraus geht jedoch oft noch keine direkte Kausalität hervor. Komplexere Ansätze, wie zum Beispiel die SHAP-Value-Analyse, die einen in der Spieltheorie fundierten Ansatz nutzt, können darüber hinaus auch Rückschlüsse auf die Kausalität der Entscheidungen geben.[19] Daher wurde hier bei den Antwortmöglichkeiten eine weitere Kategorie eingeführt.

19 Mangalathu, Sujith, Seong-Hoon Hwang, and Jong-Su Jeon. "Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach." Engineering Structures 219: 110927. 2020.

BIASED TRAINING DATA 16.4 %

How is the use of biased training data prevented?

Answer	Response Factor
No standardized process	0
Basic statistical analysis of the training data	1
Analysis of training data using specialized metrics (e.g. fairness metrics)	2

A preliminary review of the training data helps to prevent learning bias in the AI model. In order to recognize whether the training data form a fair database, different methods can be used. On the one hand, basic statistical analyses help to uncover correlations that point to distorted or biased data. Recognized *fairness metrics* analyze whether certain subgroups show systematic advantages or disadvantages in the available data, and help to make the results of the analyses comparable across several use cases. Since they are based on scientifically recognized methods, they also reduce the risk of the data analysis being (indirectly) influenced by potential bias of *data scientists* or the composition of the team.[20]

A solid and unbiased database is the only way to create a robust, unbiased AI model. This question was therefore also assigned a relatively high weighting of 16.4 %.

VOREINGENOMMENE TRAININGSDATEN 16,4 %

Wie wird verhindert, dass voreingenommene Trainingsdaten verwendet werden?

Antwort	Antwort-Faktor
Kein standardisierter Prozess	0
Grundlegende statistische Analysen der Trainingsdaten	1
Analyse der Trainingsdaten anhand spezialisierter Metriken (z. B. Fairness-Metriken)	2

Eine vorausgehende Überprüfung der Trainingsdaten hilft dabei, dem Erlernen von Voreingenommenheit im KI-Modell vorzubeugen. Um zu erkennen, ob die Trainingsdaten eine faire Datenbasis bilden, können verschiedene Methoden zum Einsatz kommen. Hier helfen grundlegende statistische Analysen dabei, Zusammenhänge aufzudecken, die auf verzerrte oder vorurteilsbelastete Daten hinweisen. Anerkannte *Fairness-Metriken* analysieren, ob bestimmte Untergruppen systematische Vor- oder Nachteile in den vorliegenden Daten aufweisen und helfen dabei, die Ergebnisse der Analysen über mehrere Anwendungsfälle hinweg vergleichbar zu machen. Da sie auf wissenschaftlich anerkannten Methoden basieren, ist zusätzlich die Gefahr geringer, dass die Datenanalyse durch potenzielle Voreingenommenheit von *Data Scientists* oder die Zusammensetzung des Teams (indirekt) beeinflusst wird.[20]

Eine solide und vorurteilsfreie Datenbasis ist die einzige Möglichkeit, ein robustes, vorurteilsfreies KI-Modell zu erstellen, dadurch wurde diese Frage mit einem Anteil von 16,4 % ebenfalls relativ hoch gewichtet.

20 Garg, Pratyush, John Villasenor, and Virginia Foggo. "Fairness Metrics: A Comparative Analysis." arXiv preprint arXiv:2001.07864. 2020.

AMOUNT OF TRAINING DATA 13.7 %

How is it ensured that enough training data is available?

Answer	Response Factor
Trial and error	0
Pre-analysis of the training data	1
Detailed analysis of the classifications and balancing of the possible imbalance by a balancing method (e.g. over/under-sampling, clustering techniques, weight balancing)	2

The amount and balance of training data have a direct impact on the complexity and significance of the insights that can be derived from the data. There is also a direct influence on how representative the training data is. It is essential to check whether there is enough training data in prior analyses. It may be necessary to balance the data to achieve a balanced data set in all classes. In this way, it can be ensured that a sufficiently large and balanced data set is available for a robust model, but the optimization process is already started regardless of this. The resulting *trial-and-error* approach will usually also come to a conclusion in the course of development that more training data is needed. If this is known beforehand, however, greater attention can be paid to other measures to increase robustness during development. In the case that the amount of training data is so small that certain regularities are not represented, this can lead to the lack of robustness only being discovered during testing with further data. Likewise, a data set that has different amounts of data in the individual classifications can lead to the model making biased decisions. To avoid the expense of the resulting fundamental changes, a weighting of 13.7 % was chosen here.

MENGE DER TRAININGSDATEN 13,7 %

Wie wird sichergestellt, dass genug Trainingsdaten vorhanden sind?

Antwort	Antwort-Faktor
Trial and Error	0
Voranalyse der Trainingsdaten	1
Ausführliche Analyse der Klassifizierungen und Ausgleich der möglichen Imbalance durch eine Balancing-Methode (z. B. Over-/Under-Sampling, Clustering-Techniken, Weight-Balancing)	2

Die Menge und die Balance der Trainingsdaten haben einen direkten Einfluss auf die Komplexität und Signifikanz der Erkenntnisse, die aus den Daten abgeleitet werden können. Zudem gibt es einen direkten Einfluss, wie repräsentativ die Trainingsdaten sind. Ob genügend Trainingsdaten vorhanden sind, sollte in vorausgehenden Analysen überprüft werden. Unter Umständen muss im Anschluss noch ein Balancing durchgeführt werden, um eine ausgewogene Datenlage in allen Klassen zu erreichen. So kann erzielt werden, dass ein ausreichend großes und ausbalanciertes Datenset für ein robustes Modell vorliegt, ungeachtet dessen aber schon der Optimierungsprozess gestartet wird. Der so entstehende *Trial-and-Error*-Ansatz wird im Laufe der Entwicklung meist ebenfalls zu dem Ergebnis kommen, dass mehr Trainingsdaten benötigt werden. Wenn dies aber schon vorher bekannt ist, kann während der Entwicklung ein größeres Augenmerk auf andere Maßnahmen zur Steigerung der Robustheit gelegt werden. Falls die Menge der Trainingsdaten so klein gewählt ist, dass bestimmte Gesetzmäßigkeiten darin nicht abgebildet sind, kann das dazu führen, dass die fehlende Robustheit erst während der Erprobung mit weiteren Daten festgestellt wird. Ebenfalls kann ein Datenset, welches unterschiedlich viele Daten in den einzelnen Klassifikationen hat, dazu führen, dass das Modell voreingenommene Entscheidungen trifft. Um den Aufwand der daraus folgenden grundlegenden Änderungen zu verhindern, wurde hier eine Gewichtung von 13,7 % gewählt.

OUTPUT DATA 13.7 %

Are the output data limited or unpredictable?

Answer	Response Factor
Nearly unlimited (dynamically generated text, audio, images)	-1
Large number of output options (large number of classes, large possible value range)	0
Restricted (classification with a small number of classes, regression in a certain range of values)	1

AUSGABEDATEN 13,7 %

Sind die Ausgabedaten eingeschränkt oder unvorhersehbar?

Antwort	Antwort-Faktor
Quasi unbegrenzt (dynamisch generierter Text, Audio, Bilder)	-1
Große Anzahl an Ausgabemöglichkeiten (große Klassenanzahl, großer möglicher Wertebereich)	0
Eingeschränkt (Klassifizierung mit geringer Klassenanzahl, Regression in bestimmtem Wertebereich)	1

The type of output data of the AI model influences its complexity. As the complexity of the model increases, the amount of data required for a robust model increases disproportionately and it becomes increasingly difficult to understand decisions. Models with more complex input data receive a lower rating. With dynamic output options such as freely generated text, traceability suffers due to the unlimited number of potential outputs. As the associated effects on robustness are particularly prominent, this response option was given a negative *response score*. However, since even use cases with highly complex output options can be implemented robustly, this question should not be weighted too heavily. In coordination with the other questions, this resulted in a score of 13.7 %.

Die Art der Ausgabedaten des KI-Modells beeinflusst dessen Komplexität. Da mit steigender Komplexität des Modells, die Menge der für ein robustes Modell benötigten Daten überproportional steigt und es zunehmend schwieriger wird, Entscheidungen nachzuvollziehen, erhalten Modelle mit komplexeren Eingabedaten eine geringere Bewertung. Gerade bei dynamischen Ausgabemöglichkeiten, wie frei generiertem Text, leidet die Nachvollziehbarkeit durch die unbegrenzte Anzahl an potenziellen Ausgaben. Da die damit verbundenen Auswirkungen auf die Robustheit besonders hervorstechen, wurde diese Antwortmöglichkeit mit einem negativen *Antwort-Score* versehen. Da aber auch Anwendungsfälle mit hochkomplexen Ausgabemöglichkeiten robust umgesetzt werden können, sollte diese Frage nicht mit einer zu hohen Gewichtung einfließen. In Abstimmung mit den anderen Fragen hat sich so eine Bewertung von 13,7 % gebildet.

DATA AUGMENTATION 13.7 %

Is the training data enhanced with known transformations or natural perturbations to imitate influences such as rain, noise, or fog?

Answer	Response Factor
Yes	1
No	0

DATEN-AUGMENTATION 13,7 %

Werden die Trainingsdaten mit bekannten Transformationen oder Natural Perturbations erweitert, um Einflüsse wie Regen, Rauschen oder Nebel nachzuahmen?

Antwort	Antwort-Faktor
Ja	1
Nein	0

In order for AI models to make reliable predictions even under adverse external influences, these scenarios must be mapped in the training data. Normally, training data used for this purpose has actually been recorded under the target circumstances, for example during rainfall (for image recognition) or with loud noises (for speech recognition). Alternatively, existing training data can be used that is then transformed to imitate an influencing factor like rain. By using such transformations, the amount of usable training data can be multiplied. For example, to imitate rain on an image, the image is darkened, noise is added, and the contrast and color intensity is reduced. In addition to clearly defined transformations for known phenomena, stochastic approaches can also be chosen to generate a new transformation function (for example with a CycleGan – Generative Adversarial Network [21]).

Damit KI-Modelle auch unter widrigen äußeren Einflüssen zuverlässige Vorhersagen treffen können, ist es notwendig, dass diese Szenarien in den Trainingsdaten abgebildet werden. Normalerweise nutzt man dafür Trainingsdaten, die tatsächlich unter Umständen wie Regen (für Bilderkennung) oder bei lautem Rauschen (für Spracherkennung) aufgenommen wurden. Zusätzlich können auch bestehende Trainingsdaten verwendet werden, die mit einer Transformation bearbeitet werden, um einen Einfluss wie Regen nachzuahmen. Durch solche Transformationen kann die Menge der nutzbaren Trainingsdaten vervielfacht werden. Um beispielsweise Regen auf einem Bild nachzuahmen, wird das Bild verdunkelt, ein Rauschen hinzugefügt und der Kontrast sowie die Farbintensität reduziert. Neben klar definierten Transformationen für bekannte Phänomene können auch stochastische Ansätze gewählt werden, um so (zum Beispiel mit einem CycleGan - Generative Adversarial Network [21]) eine neue Transformationsfunktion zu generieren.

INPUT DATA  11.0 %

What kind of input data will the model process?

Answer	Response Factor
Complex, unstructured data (text, images, audio, video)	0
Structured (tabular) data	1

EINGABEDATEN 11,0 %

Welche Art Eingabedaten wird das Modell verarbeiten?

Antwort	Antwort-Faktor
Komplexe, unstrukturierte Daten (Text, Bild, Audio, Video)	0
Strukturierte (tabellarische) Daten	1

The effect of complex input data is comparable to the effect of complex output data (see the discussion in the previous section). With the help of current analysis methods, such as the *SHAP Values Analysis*, it is often possible to understand which aspect of an input is responsible for a certain output, therefore

Der Effekt von komplexen Eingabedaten ist vergleichbar mit dem Effekt von komplexen Ausgabedaten (vgl. Abhandlung im vorherigen Abschnitt). Mithilfe von aktuellen Analysemethoden kann oft nachvollzogen werden, welcher Aspekt einer Eingabe für eine bestimmte Ausgabe verantwortlich ist und somit können komplexe Modelle besser verstanden werden, wie zum Beispiel die

21 Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

making it easier to understand complex models. While this makes it easier to understand complex input data, it still cannot fully explain the behavior. Therefore, the weighting of the input data has been set slightly below the weighting of the output data.

SHAP-Values-Analyse. Dies erleichtert zwar das Verständnis komplexer Eingabedaten, kann aber trotzdem das Verhalten nicht vollständig erklären. Deswegen wurde die Gewichtung der Eingabedaten leicht unter der Gewichtung der Ausgabedaten angesetzt.

RE-TRAINING WITH PRODUCTION DATA — 5.5 %

Should the model be further developed using production data (so-called re-training)?

Answer	Response Factor
Yes	-1
No	0

Especially with interactive AI models, production data is often used to further develop AI models during their runtime (so-called *re-training*). Since this means that the training dataset is continuously changed, the effort required to keep it representative and free of distortion increases. However, if the dynamic database is handled responsibly and the model behavior is validated before the transition from the training phase to the productive phase, the robustness will only be minimally impacted. For this reason, a low weighting value of 5.5 % and negative factor were chosen here. This is, however, fully compensated for when the corresponding measures are implemented.

RE-TRAINING MIT PRODUKTIONSDATEN — 5,5 %

Soll das Modell mittels Produktionsdaten weiterentwickelt werden (sog. Re-Training)?

Antwort	Antwort-Faktor
Ja	-1
Nein	0

Gerade bei interaktiven KI-Modellen werden Produktionsdaten oft dafür verwendet, um KI-Modelle während ihrer Laufzeit weiterzuentwickeln (sog. *Re-Training*). Da dadurch die Trainingsdatenbasis kontinuierlich geändert wird, steigt der Aufwand, um diese trotzdem repräsentativ und frei von Verzerrungen zu halten. Bei einem verantwortungsvollen Umgang mit der dynamischen Datenbasis und bei der Validierung des Modellverhaltens vor dem Übergang aus der Trainingsphase in die Produktivphase, ist die Robustheit aber kaum negativ beeinflusst. Deshalb wurde hier ein niedriger Gewichtungswert von 5,5 % gewählt, der negativ angerechnet und bei einer Umsetzung der entsprechenden Maßnahmen allerdings vollständig kompensiert wird.

PREVENTION OF UNETHICAL BEHAVIOR THROUGH RE-TRAINING — 6.8 %

How will the learning of unethical behavior through re-training be prevented?

Answer	Response Factor
No tests planned/implemented	0
Quality checks using test data	1
Detailed statistical analyses	2

In order to counteract the negative effects of a continuous further development of the AI model by production data, it should always be ensured that the behavior of the model has not deteriorated before the model is transferred to productive operation. If the database is large enough, a standardized set of test data can be created for this purpose. This test data set can then be used for verification purposes before each introduction of a new model version. However, it would be more accurate to perform the verification with the help of extensive statistical tests, as this way it is also possible to verify conditions that were not included in the standardized test data set from the beginning (e.g. current trends). The weighting was chosen in such a way that, if the appropriate measures are implemented, the negative influence of the previous question will be compensated for.

VORBEUGEN VON UNETHISCHEM VERHALTEN DURCH RE-TRAINING — 6,8 %

Wie soll verhindert werden, dass durch Re-Training unethisches Verhalten erlernt wird?

Antwort	Antwort-Faktor
Keine Prüfungen geplant/umgesetzt	0
Qualitätsprüfungen mittels Testdaten	1
Ausführliche statistische Analysen	2

Um den negativen Effekten einer kontinuierlichen Weiterentwicklung des KI-Modells durch Produktionsdaten entgegenzuwirken, sollte vor der Überführung des Modells in den Produktivbetrieb immer sichergestellt werden, dass das Verhalten des Modells nicht verschlechtert wurde. Bei einer ausreichend großen Datenbasis kann dafür ein standardisierter Satz von Testdaten erstellt werden. Dieser Testdatensatz kann dann vor jeder Einführung einer neuen Modellversion für die Verifizierung genutzt werden. Genauer wäre es allerdings, die Überprüfung mithilfe von ausgiebigen statistischen Tests durchzuführen, da so auch Gegebenheiten überprüft werden können, die nicht von Anfang an im standardisierten Testdatensatz enthalten waren, weil sie beispielsweise auf Trends zurückzuführen sind. Die Gewichtung wurde so gewählt, dass bei einer Umsetzung der entsprechenden Maßnahmen der negative Einfluss der vorherigen Frage kompensiert wird.



Summary

During the development of the Robust AI Assessment, we created a framework which has the goal of making AI projects more fault-tolerant and robust. With the help of expert interviews, research of scientific publications and initial tests with current development projects, we have formed an assessment logic that allows us to make a robust assessment of how robust an AI model should be and how robust it actually is or will be implemented.

The self-assessment concept has already been tested in 12 internal projects. The result values of *Actual Robustness* and *Required Robustness* did not show any unexpected discrepancies.

The interactions with AI experts during the expert interviews and during the project tests also showed that the separate consideration of robustness contributes to the increase of the model quality. This put a clearer focus on the topic of robustness, which was previously mainly subordinated to the item "prediction quality". Predictions can be very high-quality, but at the same time they can be very unstable, which quickly impacts results. Having a clear separation between these two topics helps to increase quality and fault tolerance, which leads to the development of better AI models. In the interviews with experts and in test runs, these advantages were explicitly and positively emphasized several times.

Zusammenfassung

Während der Entwicklung des Robust AI Assessments ist ein Rahmenwerk entstanden, welches das Ziel verfolgt, KI-Projekte fehlertoleranter und robuster zu gestalten. Mithilfe von Interviews mit Expertinnen und Experten, Recherche wissenschaftlicher Publikationen und ersten Tests mit aktuellen Entwicklungsprojekten haben wir eine Bewertungslogik geformt, die es uns erlaubt, eine belastbare Einschätzung darüber zu treffen, wie robust ein KI-Modell sein sollte und wie robust es tatsächlich umgesetzt ist oder werden wird.

Das Self-Assessment-Konzept wurde bereits in 12 internen Projekten erprobt. Die Ergebniswerte der *Actual Robustness* und *Required Robustness* zeigten dabei keine unerwarteten Ausreißer.

Die Interaktion mit KI-Expertinnen und -Experten während der Interviews und Projekttests hat zusätzlich gezeigt, dass die gesonderte Betrachtung von Robustheit zur Steigerung der Modellqualität beiträgt. Das Thema Robustheit, welches vorher hauptsächlich dem Punkt „Vorhersagequalität“ untergeordnet war, wurde so mehr in den Mittelpunkt gestellt. Vorhersagen können zwar qualitativ sehr gut sein, aber gleichzeitig sehr instabil, sodass das Ergebnis schnell beeinflusst wird. Die klare Trennung zwischen diesen beiden Themen hilft dabei, die Qualität und Fehlertoleranz zu steigern, was insgesamt zur Entwicklung besserer KI-Modelle führt. In den Interviews mit den Expertinnen und Experten sowie den testweisen Durchführungen wurden diese Vorteile mehrfach explizit positiv hervorgehoben.



Critical Appraisal

Kritische Würdigung

The development of the Robust AI Assessment is based on expert interviews and scientific publications, but has not yet been scientifically independently verified *(peer review)*. It is therefore possible that there is further potential for improvement, especially with regard to the weighting of questions and domains. However, starting the procedure with an initial assessment from the expert interviews and then refining it through tests with internal projects has shown that the results correspond to the qualitative estimates of the experts and are within a plausible range without significant discrepancies.

Die Entwicklung des Robust AI Assessments basiert auf Interviews mit Expertinnen und Experten sowie wissenschaftlichen Publikationen, wurde aber bisher nicht wissenschaftlich unabhängig überprüft *(Peer-Review)*. Es ist daher möglich, dass gerade hinsichtlich der Gewichtung der Fragen und der Domänen weiteres Verbesserungspotenzial vorhanden ist. Das Vorgehen mit einer initialen Abschätzung aus den Interviews zu starten und diese dann durch Tests mit internen Projekten zu verfeinern, hat jedoch gezeigt, dass die Ergebnisse mit den qualitativen Schätzungen der Sachkundigen übereinstimmen und sich in einem plausiblen Rahmen ohne nennenswerte Ausreißer bewegen.



Outlook

The research field of data analysis and AI is continuously being developed. As already mentioned in several previous chapters, there are many new findings, especially in the context of developing new measures to increase robustness, which have been tested in research and are now also being applied by companies. In order to keep up with the current state of research, the evaluation model will have to be extended by the new findings and methods.

As mentioned above, the Robust AI Assessment focuses on the social responsibility of digital companies during the development of new technologies. The self-assessment concept is part of a larger framework that goes beyond robustness to address overarching ethical issues. This framework for assessing whether AI projects act in accordance with the DT Guidelines for Digital Ethics is currently being developed. In combination, it will allow us to make a holistic assessment of the extent to which our ethical and moral demands on new AI developments are being implemented.

Ausblick

Das Forschungsfeld Datenanalyse und KI wird kontinuierlich weiterentwickelt. Wie bereits in einigen vorherigen Kapiteln angeführt, gibt es gerade im Rahmen der Entwicklung neuer Maßnahmen zur Steigerung der Robustheit viele neue Erkenntnisse, die in der Forschung erprobt wurden und nun auch durch Unternehmen angewendet werden. Um mit dem jeweils aktuellen Forschungsstand mitzuhalten, wird das Bewertungsmodell um die neuen Erkenntnisse und Methoden erweitert werden müssen.

Wie bereits Anfangs erwähnt, fokussiert sich das Robust AI Assessment darauf, die soziale Verantwortung digitaler Unternehmen während der Entwicklung neuer Technologien zu betrachten. Das Self-Assessment-Konzept ist dabei Teil eines größeren Frameworks, das sich über die Robustheit hinaus mit übergreifenden ethischen Fragestellungen beschäftigt. Dieses Framework zur Einschätzung, ob KI-Projekte entlang der Telekom Leitlinien für digitale Ethik handeln, wird aktuell entwickelt. In Kombination kann damit dann eine ganzheitliche Abschätzung abgegeben werden, inwiefern unsere ethischen und moralischen Ansprüche an neue KI-Entwicklungen umgesetzt werden.



Appendix: Calculation example and interpretation

The calculation of the Robust AI System consists of the comparison of the Required Robustness Score and the Actual Robustness Score. Both scores will reach a value between 0–5 which, in its comparison, allows an interpretation as to how safe the AI is in its current state and how high its robustness should be. Here we use a fictitious example to make the calculation comprehensible.

8.1. CALCULATION OF THE REQUIRED ROBUSTNESS SCORE

The calculation of the Required Robustness Score results from the weighting of the questions, which are given as percentages and multiplied by the score given to the questions.

E.G. QUESTION 1: WORST-CASE-SCENARIO ⚖ 16.7 %

What kind of worst-case-scenario could theoretically be triggered by the model making a wrong decision?

Answer	Response Factor
High impact: The emergence of the risk forces the company to change its objectives or strategy in the short-term. Example: A failure leads to network instability on a large scale.	2
Medium impact: The emergence of the risk requires medium-term changes to the company's objectives or strategy. Example: The error leads to reputational damage, which receives attention from leading media sources.	1
Trivial impact: No effect on the company value. Example: Internal processes cannot take place at the usual speed.	0

Anhang: Rechenbeispiel und Inter-pretation

Die Berechnung des Robust AI Systems besteht aus der Gegenüberstellung des Required Robustness Scores und des Actual Robustness Scores. Beide Scores werden einen Wert zwischen 0 und 5 erreichen, der in seiner Gegenüberstellung eine Interpretation erlaubt, wie sicher die KI in ihrem aktuellen Zustand ist und wie hoch ihre Robustheit sein sollte. Wir nutzen hier ein fiktives Beispiel, um die Berechnung nachvollziehbar zu machen.

8.1. BERECHNUNG DES REQUIRED ROBUSTNESS SCORES

Die Berechnung des Required Robustness Scores entsteht aus der Gewichtung der Fragen, die in Prozenten angegeben sind und werden mit der Punktzahl, die für die Fragen vergeben wurden, multipliziert.

Z. B. FRAGE 1: WORST-CASE-SZENARIO ⚖ 16,7 %

Welche Art von Worst-Case-Szenario könnte theoretisch durch eine Fehlentscheidung des Modells ausgelöst werden?

Antwort	Antwort-Faktor
Hoher Einfluss: Das Auftreten des Risikos zwingt das Unternehmen, seine Ziele oder Strategie kurzfristig zu ändern. Beispiel: Ein Ausfall führt zu Netzinstabilität im großen Maß.	2
Mittlerer Einfluss: Das Auftreten des Risikos erfordert mittelfristige Änderungen der Ziele oder der Strategie des Unternehmens. Beispiel: Der Fehler führt zu einem Reputationsschaden, welcher von führenden Medien aufgegriffen wird.	1
Trivialer Einfluss: Keine Auswirkung auf den Unternehmenswert. Beispiel: Interne Prozesse können nicht in gewohnter Geschwindigkeit erfolgen.	0

⚖ = Weighting of questions / Fragengewichtung

Question 1 / Frage 1

Weighting of the question: /
Gewichtung der Frage: 0,167

$2 * 0{,}167 = 0{,}334$

Question 2 / Frage 2

Weighting of the question: /
Gewichtung der Frage:

$1 * 0{,}14 = 0{,}139$

Question 3 / Frage 3

...
...

$$Conversion\ factor = \frac{Maximum\ target\ value}{\sum_{i=1}^{n} Answer_i * Weighting_i}$$

$$Conversion\ factor_{\substack{Required\\Robustness}} = \frac{5}{0{,}334 + 0{,}139 + 0{,}194 + 0{,}028 + 0{,}069 + 0{,}097 + 0{,}139 + 0{,}111 + 0{,}111 + 0{,}056 + 0{,}112} = 3{,}59$$

The conversion factor: 3.59 is used to calculate the score. /
Der Umrechnungsfaktor: 3,59 wird genutzt, um den Score auszurechnen.

Fictitious example 1 / Fiktives Beispiel 1

Addition of the evaluation scores that are first multiplied by the question weighting: /
Addition der Bewertungspunktzahlen, die mit der Fragengewichtung zuerst multipliziert werden:

$$Score_{Required\ Robustness} = Conversion\ factor_{\substack{Required\\Robustness}} * \sum_{i=1}^{n} Answer_i * Weighting_i$$

$$Score_{Required\ Robustness} = 3{,}59 * \Big((1 * 0{,}167) + (0 * 0{,}139) + (1 * 0.097) + (1 * 0.056) + (1 * 0{,}028) + (0 * 0{,}111) + (1 * 0{,}111) + (2 * 0{,}097) + (1 * 0{,}069) + (1 * 0{,}069) + (2 * 0{,}056) \Big) = 3{,}24$$

The Required Robustness Score is 3.24. /
Der Required Robustness Score liegt bei 3,24.

On the scale, the Actual Robustness Score should now reach at least the value of 3.24 to ensure that the robustness requirements of the AI are met.
The calculation of the Actual Robustness Score varies only in that the three sub-categories

- Robustness of the process around the AI used
- Robustness of the AI model used and
- Robustness of the data model used

are averaged for the overall result of the Actual Robustness Score

Auf der Skala sollte nun der Actual Robustness Score mindestens den Wert von 3,24 erreichen, um sicherstellen zu können, dass die Anforderungen an die Robustheit der KI erfüllt werden.
Die Berechnung des Actual Robustness Scores variiert nur dahingehend, dass die drei Unterkategorien

- Robustheit des Prozesses rund um die eingesetzte KI,
- Robustheit des eingesetzten KI-Modells und
- Robustheit des verwendeten Datenmodells

für das Gesamtergebnis des Actual Robustness Scores gemittelt werden.

$$Score_{Actual\ Robustness} = \frac{Score_{Process} + Score_{AI\ Model} + Score_{Data}}{3}$$

Calculate the conversion factor for the three circle segments as shown above: /
Umrechnungsfaktor für die drei Kreisabschnitte berechnen, wie oben gezeigt:
Conversion factor Actual Robustness Process: /
Umrechnungsfaktor Actual Robustness Prozess: 2,67
Conversion factor Actual Robustness KI Model: /
Umrechnungsfaktor Actual Robustness KI-Modell: 4,35
Conversion factor Actual Robustness Data: /
Umrechnungsfaktor Actual Robustness Data: 3,65

Example values with fictitious answers: /
Beispielwerte mit fiktiven Antworten:

$Score_{Process} = 1{,}3 * 2{,}67 = 3{,}47$

$Score_{AI\ Model} = 0{,}7 * 4{,}35 = 3{,}05$

$Score_{Data} = 0{,}9 * 3{,}65 = 3{,}28$

$$Score_{Actual\ Robustness} = \frac{3{,}47 + 3{,}05 + 3{,}28}{3} = 3{,}27$$

The Actual Robust Score is 3.27 points. /
Der Actual Robust Score beträgt 3,27 Punkte.

If you compare the two scores with each other, you will find that the actual robustness slightly exceeds the required robustness, revealing that the solution is sufficiently robust.

Vergleicht man die beiden Scores miteinander, stellt man fest, dass die tatsächliche Robustheit die notwendige leicht übersteigt und die Lösung somit ausreichend robust ist.

9

Imprint

Impressum

Further information and contacts: /
Weitere Informationen und Kontakte:

Deutsche Telekom AG
Group Compliance Management
Friedrich-Ebert-Allee 140
53113 Bonn

Digital.Ethics@Telekom.de



https://www.telekom.com/de/konzern/digitale-verantwortung/ethische-ki-leitlinien-der-telekom

Bonn, 31.03.2021

Authors: /
Autor*innen:

Manuela Mackert
Manuela.Mackert@Telekom.de

Manuel Mikoleit
M.Mikoleit@Telekom.de

Editorial Design & Realization: /
Gestaltung & Realisation:

Sensity sàrl
www.sensity.eu



